Improved local polynomial estimation in nonparametric time series regression

D i s s e r t a t i o n

zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)

vorgelegt dem Rat der Fakultät für Mathematik und Informatik
der Friedrich-Schiller-Universität Jena

von Dipl.-Math. Juliane Geller
geboren am 17.03.1987 **in** Jena

1. Gutachter: Prof. Dr. Michael H. Neumann (Friedrich-Schiller-Universität Jena)

2. Gutachter: Prof. Dr. Gustau Camps-Valls (Universitat de València)

Tag der öffentlichen Verteidigung: 20. Oktober 2016

Zusammenfassung

In der vorliegenden Arbeit wird eine Modifikation des lokal polynomialen Schätzers im nichtparametrischen Regressionsmodell für abhängige Fehlerdaten vorgestellt. Diese berücksichtigt die Autokorrelation der Beobachtungsfehler und liefert damit effizientere Schätzungen als der konventionelle lokal polynomiale Schätzer. Wir verallgemeinern damit die Resultate von Xiao et al. (2003) und Su & Ullah (2006), welche effizientere Schätzer für Fehlerdaten mit funktionaler Abhängigkeitsstruktur entwickelt haben. Im Gegensatz zu diesen Arbeiten setzen wir lediglich voraus, dass die Abhängigkeit der Fehler schnell genug gegen null konvergiert. Mathematisch wird diese Voraussetzung durch einen stark mischenden Fehlerprozesses modelliert. Die meisten gängigen autoregressiven Prozesse, wie etwa ARMA - und viele ARCH - Modelle, sind stark mischend.

Die Modifizierung des Schätzers beruht auf einer sogenannten *prewhitening* - Transformation der Daten, bei welcher der Fehlerprozess durch einen AR - Prozess approximiert wird. Die Fehlerterme werden dabei durch die Residuen einer vorausgehenden lokal polynomialen Regression geschätzt. Wir leiten die asymptotische Verteilung des resultierenden Schätzers her und zeigen damit, dass dieser asymptotisch effizienter als der konventionelle Schätzer ist. Dieses Resultat ist insofern überraschend, als dass die tatsächliche Kovarianzstruktur des Fehlerprozesses stark von der implizit angenommenen AR - Struktur abweichen kann.

In den meisten Standardmodellen wird ein varianzhomogener Fehlerprozess angenommen. In vielen Anwendungen aus dem Finanzbereich und der Ökologie ist diese Forderung jedoch verletzt. Um auch Fehlerprozesse mit inhomogener Varianz zu berücksichtigen, erweitern wir unser Schätzverfahren auf ein heteroskedastisches Regressionsmodell. Wir zeigen, dass unser Schätzer auch unter Heteroskedastizität asymptotisch effizientere Schätzungen liefert als der konventionelle. In diesem Zusammenhang geben wir gleichmäßige Konvergenzraten für den Schätzer der Varianzfunktion an. Damit verallgemeinern wir die Arbeiten von Ruppert et al. (1997) und Fan & Yao (1998), in welchen punktweise Konvergenzraten bewiesen wurden.

Eine Simulationsstudie soll unsere asymptotischen Aussagen veranschaulichen. Wir gehen zunächst auf die Implementierung unseres Schätzers ein und wenden ihn dann auf mehrere Regressionsmodelle mit linearen und nichtlinearen Fehlerprozessen an. Die Ergebnisse legen nahe, dass unsere Schätzmethode selbst in mittelgroßen Datensätzen zum Teil erhebliche Effizienzverbesserungen liefert.

Wir schließen die Arbeit mit der Diskussion von möglichen Verallgemeinerungen der hier betrachteten Modelle und Schätzmethoden ab.

Abstract

We propose a modification of local polynomial estimation in nonparametric regression that improves the efficiency over the conventional estimator when the observation errors are autocorrelated. This generalizes the works of Xiao et al. (2003) and Su & Ullah (2006), who considered error processes with a certain functional autocorrelation structure. In contrast to that, we do not suppose any covariance structure on the error process. We only need the dependence between two observation errors to vanish sufficiently fast as their corresponding time lag increases. More precisely, we suppose the error process to be strongly mixing. This assumption covers a wide range of processes, including the popular autoregressive moving average (ARMA) and autoregressive conditional heteroscedasticity (ARCH) models.

The modified procedure is based on a prewhitening transformation of the data. More precisely, we approximate the error process with an AR process. In order to do so, we first estimate the observation error via preliminary local polynomial smoothing. Deriving its asymptotic distribution, we show that the resulting estimator is more efficient than the conventional one. This result is quite surprising, since the actual error covariance structure can be quite far from the implicitly supposed AR correlation structure.

Conventionally, the observation error is supposed to possess a constant variance. However, this assumption is strongly violated in a wide set of applications in finance and ecology. We address this problem mathematically by extending the prewhitening transformation to a heteroscedastic regression model. As for the homoscedastic setting, we prove that the resulting estimator is asymptotically more efficient than the conventional one. In this context, we obtain uniform convergence rates for the estimator of the variance function. This extends the results of Ruppert et al. (1997) and Fan & Yao (1998), who provided pointwise convergence rates.

A simulation study is performed in order to illustrate the finite sample performance of the proposed estimator and give empirical evidence of the asymptotic results. We first address the implementation of our estimator, before we apply it to various special cases of linear and nonlinear strongly mixing error processes in both the homoscedastic and the heteroscedastic regression model. The obtained results give evidence that significant efficiency gains can be achieved even for a moderate number of observations.

We conclude this thesis by discussing some possible generalizations that could be the object of future research.

Acknowledgments

Pursuing a Ph.D. project can be a rather challenging experience at sometimes and I want to take this opportunity to thank all the people that kindly supported and encouraged me on my way.

I gratefully acknowledge the support of my supervisors Prof. Dr. Michael H. Neumann and Dr. Jens Schumacher, for providing me with the assistance and expertise that I needed. Without their guidance and constant feedback I would not have been able to pursue this thesis to the end.

I further want to express my deep gratitude to Prof. Dr. Gustau Camps-Valls, who supervised me during my time in his lab in Valencia. I thank him for this incredible research stay and for showing me that one can conduct very good research in a relaxed and joyful lab atmosphere.

This thesis was conducted in the framework of the IMPRS grad school, that I would like to thank for both the financial and ideal support. This grad school provided me with courses that helped me to develop my skills as a researcher and gave me the great opportunity of a foreign research stay. I would also like to thank my MPI colleagues Prof. Dr. Markus Reichstein, Dr. Miguel Mahecha and Dr. Martin Jung for their big interest in my research.

I would not have been able to complete this thesis without the financial support of Prof. Dr. Joachim Giesen, who offered me a six month position as a research assistant in his group. I am most grateful for this opportunity.

Special thanks are also given to Prof. Dr. Ingo Althöfer, who was the one that pointed out this Ph.D. project to me.

Concerning all the technical problems that I had during my research, I would like to acknowledge the help of Erik Braun and Uwe Richter from the *Rechenzentrum*.

My working days were filled with pleasure thanks to my wonderful colleagues: Astrid Jeschag, Carme Zaragoza, Christian Ludwig, Clara Petersen, Jorge Vicent Servera, Marlis Bärthel, Martin Semisch, Raúl Santos-Rodríguez, Stefan Ankirchner, and Stefan Rosemann are just a part of the people that spent so many nice breaks with me. I am thankful to Christopher Schneider for always finding a fast and good solution to the numerous LaTeX concerns that came up while writing this thesis. I am deeply grateful to my colleague Nadine Müller, for just being the best person that one can possibly share an office room with.

I am strongly indebted to every single one of my friends who incented me to strive towards the completion of this thesis. I also want to express my gratitude to my family in law - Kerstin Undeutsch and Emma, Finja, and Peter Perschke, who have been so supportive and caring.

Without the great love and understanding of my parents Dörthe and Klaus Geller, I am not sure if I would have done all the meaningful things in my life. I will miss my Mom dropping by the office to provide us with food or anything else that came into her mind.

Finally, words cannot express how thankful I am to my beloved partner Paula Perschke who suffered with me in all the stressful moments during the final year of my Ph.D. Her love and understanding form big parts of my inner happiness.

Jena, 2016-07-20 Juliane Geller

Contents

List of illustrations

Figures

Tables

Chapter 1

Introduction

Regression analysis forms one of the key parts of mathematical statistics. The aim of this statistical tool is to estimate the functional relationship between a dependent variable and one or more explanatory variables in the presence of observation errors. In their basic definition, most regression methods assume these errors to be independent. In practice, however, one is often confronted with the situation where the noise is no longer white, and instead contains a certain amount of "structure" in the form of autocorrelation. Examples for this phenomenon are applications from financial predictions, signal processing and system observation where the observations come from a time series, as the relationship between weather and electricity sales (Engle et al. 1986) and financial stock returns (Ferenstein & Gaşowski 2004). For such data sets, information is left out when the error dependency is ignored and one can expect a more precise estimate by including the autocorrelation in the estimation process.

For univariate linear regression models of the form

$$Y_t = \beta X_t + U_t, \qquad t = 1, \ldots, n,$$

with a stationary error process $(U_t)_{t \geq 1}$ that is autocorrelated but has zero mean, it is well-known that the generalized least squares (GLS) estimator for β improves the efficiency of the conventional ordinary least squares (OLS) estimator by including the covariance matrix $\mathrm{Cov}(U_i, U_j)$, $i, j = 1, \ldots, n$ in the estimation process. Typically, this covariance matrix is unknown and has to be estimated a priori. If the prior estimation converges sufficiently fast to the true covariance matrix, then the resulting feasible GLS estimator is asymptotically equivalent to the GLS estimator and hence more efficient than its OLS counterpart.[1]

In this thesis, we want to transfer the main underlying idea of the GLS estimator to the nonpara-

[1] See for example Aitken (1934) for more detailed information on GLS estimation.

metric regression model

$$Y_t = m(X_t) + U_t, \qquad t = 1, \ldots, n. \tag{1.1}$$

We assume the conditional mean function $m(\cdot)$ to be unknown but smooth. Our aim is to provide a more efficient estimator of this function.

If the explanatory variables are fixed, e.g. $X_t = x_t \in \mathbb{R}$ for $t = 1, \ldots, n$, then adjusting for autocorrelation brings no advantage in terms of efficiency.[2] For this reason, we consider the random design case. More precisely, we suppose that the sequence of observations originates from a stationary process $(X_t, Y_t)_{t\in\mathbb{Z}}$.

We focus on local polynomial fitting as estimation method for the conditional mean, thanks to its attractive properties including its simplicity, its minimax property, and its automatic boundary adaptation (Ruppert & Wand 1994, Fan & Gijbels 1996). In Chapter 2 we provide a brief introduction of this estimator.

In its standard formulation, the local polynomial estimator does not take into account the error correlation structure and estimates the regression function in the same way as if the observations were independent. However, recently there has been growing interest in more efficient local polynomial estimation for dependent error processes. Most research in this direction has focused on panel data with a parametric error covariance structure. Several authors have tried to account for the correlation within a specific cluster in such datasets, see for example Severini & Staniswalis (1994) and Wild & Yee (1996). However, Lin & Caroll (2000) showed that in many random effects panel data models it is better to ignore the correlation within a cluster entirely, the so-called "working independence" approach. A few years later, Caroll et al. (2004) constructed a kernel-type method that can take advantage of the correlation structure among the data. Recently, Martins-Filho & Yao (2009) and Su et al. (2013) proposed more efficient nonparametric local linear estimation procedures for a general parametric error covariance structure.

It was not clear how these findings could be extended to the regression model (1.1) until Xiao, Linton, Caroll & Mammen (2003) adopted the "prewhitening" method introduced by Ruckstuhl et al. (2000). Their approach can be seen as a natural extension of the GLS estimator to the nonparametric regression setting. We rely our research on their proposed estimation method and asymptotic results.

These authors defined a more efficient two-step procedure for the local polynomial estimator. They assumed the error process $(U_t)_{t\in\mathbb{Z}}$ to be stationary, centered, and to have an invertible linear

[2] See Xiao et al. (2003) for a discussion on adjusting for autocorrelation for fixed explanatory variables, and Vilar-Fernández & Francisco-Fernández (2002) for asymptotic results for the fixed design case.

process representation:

$$U_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}, \qquad \text{with } \varepsilon_t \text{ i.i.d., } \mathbb{E}(\varepsilon_t) = 0, \operatorname{Var}(\varepsilon_t) = \sigma_\varepsilon^2, \text{ for } t = 1, \ldots, n.$$

As for the GLS estimator, the basic idea of their estimation procedure is to "prewhiten" the data in a way, that the resulting error terms are uncorrelated: If the error terms $(U_{t-1}, U_{t-2}, \ldots)$ are observable, we can find $a(L) := \sum_{j=0}^{\infty} a_j L^j$ where L is the usual lag operator, such that $a(L)U_t = \varepsilon_t$. Local polynomial smoothing on the dataset $a(L)Y_t, t = 1, \ldots, n$ is more efficient than on the original dataset since $\operatorname{Var}(U_t) \geq \sigma_\varepsilon^2$. However, in practice the error terms are unknown. Further, only a finite order AR prewhitening is feasible. As for the feasible GLS estimator, the authors thus replaced the error terms by the residual series $(\widehat{U}_{t-1}, \ldots, \widehat{U}_{t-q})$ obtained via prior smoothing, and $a(L)$ by a corresponding finite approximation. They showed that such a replacement does not effect the first-order asymptotic efficiency of the resulting estimator, that is hence more efficient than the conventional one.

Several extensions of this estimator have been made. Linton & Mammen (2008) proposed a more efficient local polynomial estimator for the semiparametric lag model and showed that it improves the performance of the estimator introduced above in this setting. Su & Ullah (2006) derived a more efficient local polynomial estimation process for a nonparametric functional error structure.

The aim of the present thesis is to generalize these past findings and weaken the assumption of a given covariance structure for the error process. Here, we only assume that the autocorrelation between the errors vanishes as the gap between their corresponding time points increases. More precisely, we assume that the error process satisfies certain strong mixing conditions. The strong mixing property is used to describe time series that do not necessarily "fit" any specific correlation structure but are asymptotically independent. However, several time series models such as ARMA and ARCH which are often used to describe, e.g. financial data, are strongly mixing under natural conditions (Doukhan 1994, Bradley 2007). In Chapter 2, we give a brief introduction to this concept of weak dependence.

After introducing the local polynomial regression estimator for strongly mixing observations in Chapter 2, we propose a modified estimator in Chapter 3 that exploits the error-dependency. Deriving its asymptotic distribution, we prove that our proposed estimator is asymptotically more efficient than the conventional one. We apply the two-step procedure of Xiao et al. (2003) as introduced above. We shall see that their asymptotic findings remain valid in the very general strong mixing error case. This result is quite surprising, since the error process structure (if there is any) can be far away from the implicitly supposed linear AR structure.

The second milestone of this thesis is the extension of our findings to the heteroscedastic setting in Chapter 4. In our context, heteroscedasticity describes the phenomenon of an inhomogeneous variance across the error terms. More precisely, the (conditional) variance of the error term changes with the value of the design variable. This typically occurs in data sets in which there is a large disparity in the measurement accuracy of the data. In Chapter 4, we briefly introduce some of the manifold statistical applications of heteroscedasticity. In order to describe this phenomenon mathematically, we extend our regression model in Chapter 4 to

$$Y_t = m(X_t) + \sigma(X_t)U_t, \qquad t = 1, \ldots, n, \tag{1.2}$$

with $\sigma(\cdot) > 0$ an unknown but smooth variance function. We adjust the filtering transformation technique to model (1.2) and show that even in this setting we can construct an estimator that proves to be asymptotically normal and more efficient than the conventional local polynomial estimator.

Note that in order to perform a filtering transformation of the data in the heteroscedastic regression model, we additionally have to estimate the variance function. In this context we obtain a uniform bound for the variance estimator based on local polynomial smoothing on the squared residuals. This result complements the works of Fan & Yao (1998) and Ruppert et al. (1997), who provided pointwise convergence results for similar variance estimators.

In Chapter 5 we address some practical issues regarding the implementation of our estimators. In a Monte Carlo study, we apply our estimators to various scenarios of linear and nonlinear error processes with constant or point-dependent variance. The obtained results give evidence that significant efficiency gains can be achieved in finite samples.

The concluding Chapter 6 addresses potential extensions for future work on more efficient non-parametric regression.

Chapter 2

Local polynomial regression estimation for strongly mixing processes

We introduce the local polynomial estimator and its asymptotic properties as a basis for the main result of this thesis: the development and analysis of more efficient estimators in Chapters 3 and 4. In the first part of this chapter we provide a short summary of local polynomial estimation including related literature. We then establish the asymptotic normality of this estimator for strongly mixing observations in the second part of this chapter. In this context, we derive asymptotic expressions for the bias and the variance term.

Contents

2.1 Local polynomial regression estimation

Local polynomial estimation, also called local polynomial fitting or smoothing, is a nonparametric regression technique that has been around for a long time. It was first systematically studied by Stone (1977), Cleveland (1979), and Tsybakov (1986), and the asymptotic properties were established by Fan (1992, 1993), Fan & Gijbels (1992), and Ruppert & Wand (1994). A broad review of this method can be found in the monograph of Fan & Gijbels (1996). Local polynomial fitting possesses manifold attractive properties, namely its automatic adaption to the boundary of the design points (Fan & Gijbels 1992, Ruppert & Wand 1994), its minimax property (Fan 1993, Fan et al. 1997) and the availability of fast algorithms (Seifert et al. 1994). Ruppert & Wand (1994) and Fan & Gijbels (1995) also emphasized the goodness of local polynomial fitting for derivative estimation.

In this estimation method, a polynomial weighted least squares regression of order p is fitted at each point of interest x using only data from some neighborhood around x, which is determined by a bandwidth $h(n) = h_n$. Furthermore, the data is weighted by a positive real function $K(\cdot)$, that will be called the *kernel* here. This function assures that all data points outside the determined neighborhood are not taken into account. For this purpose we assume bounded support for the kernel.

Mathematically speaking, we abstract our problem as follows: recall the regression model that we introduced in the previous chapter

$$Y_t = m(X_t) + \sigma(X_t)U_t, \qquad t = 1, \dots, n,$$

with the variance function $\sigma^2(x) := \mathrm{Var}(Y|X = x) > 0$, and the error terms coming from a stationary[1] process $(U_t)_{t\in\mathbb{Z}}$ with mean zero and finite variance σ_U^2. The conditional mean function in the point $X = x$

$$m(x) := \mathbb{E}(Y|X = x)$$

is supposed to be unknown but smooth and is the object of central interest.

If $m(\cdot)$ has $p + 1$ continuous derivatives, the Taylor expansion of $m(X_t)$ around x with Lagrange form of the remainder[2] provides

$$\begin{aligned} m(X_t) = \;& m(x) + \sum_{j=1}^{p} \frac{m^{(j)}(x)}{j!} h_n^j \left(\frac{X_t - x}{h_n}\right)^j + \frac{m^{(p+1)}(x)}{(p+1)!} h_n^{p+1} \left(\frac{X_t - x}{h_n}\right)^{p+1} \\ & + \frac{m^{(p+1)}(\xi_t) - m^{(p+1)}(x)}{(p+1)!}\, h_n^{p+1} \left(\frac{X_t - x}{h_n}\right)^{p+1} \\ =:& \sum_{j=0}^{p} \beta_j(x) \left(\frac{X_t - x}{h_n}\right)^j + b_n(x) + e_n(x), \end{aligned} \tag{2.1}$$

for some real value ξ_t between x and X_t. Neglecting the terms $b_n(x)$, and $e_n(x)$, the last equation models $m(X_t)$ locally by a polynomial of order p. This suggest to estimate $\boldsymbol{\beta}(x) := \left(m(x), \dots, \frac{m^{(p)}(x)}{p!} h_n^p\right)^{\mathsf{T}}$ per sample point x by a locally weighted polynomial regression in the

[1] Throughout this thesis, when speaking of a stationary random process we refer to a weakly stationary (or covariance stationary) process, as defined for example in Priestley (1981).

[2] See for example Spivak (2006, Chapter 20).

neighborhood $[x - h_n, x + h_n]$:

$$\widehat{\boldsymbol{\beta}}(x) = \begin{pmatrix} \widehat{\beta}_0(x) \\ \vdots \\ \widehat{\beta}_p(x) \end{pmatrix} := \text{argmin}_{\beta_0,\ldots,\beta_p} \sum_{t=1}^{n} \left(Y_t - \sum_{k=0}^{p} \beta_k \left(\frac{X_t - x}{h_n} \right)^k \right)^2 \frac{1}{nh_n} K\left(\frac{X_t - x}{h_n} \right). \tag{2.2}$$

The Taylor expansion provides that $\widehat{m^{(k)}}(x) := \widehat{\beta}_k(x)\ k!\ h_n^{-k}$ is an estimator for $m^{(k)}(x)$. The whole curve $m(\cdot)$ is of course obtained by running a local polynomial regression with x varying in the corresponding estimation domain.

It is more convenient to work with matrix notation. Put $\mathbf{Y} := (Y_1, \ldots, Y_n)^\mathsf{T}$ and denote by $\mathbf{X}$ the design matrix of the problem defined in equation (2.2) and $\mathbf{W}$ the $n \times n$ diagonal matrix of weights:

$$\mathbf{X} := \begin{pmatrix} 1 & \frac{X_1 - x}{h_n} & \ldots & \left(\frac{X_1 - x}{h_n}\right)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \frac{X_n - x}{h_n} & \ldots & \left(\frac{X_n - x}{h_n}\right)^p \end{pmatrix}, \mathbf{W} := \text{diag}\left(\frac{1}{nh_n} K\left(\frac{X_1 - x}{h_n}\right),\ \ldots,\ \frac{1}{nh_n} K\left(\frac{X_n - x}{h_n}\right) \right).$$

The least squares problem defined in equation (2.2) can then be written as

$$\widehat{\beta}(x) := \text{argmin}_{\boldsymbol{\beta}} \left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right)^\mathsf{T} \mathbf{W} \left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right),$$

with $\boldsymbol{\beta} := (\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_p)^\mathsf{T}$. If the matrix $\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}$ is positive definite, the solution $\widehat{\boldsymbol{\beta}}(x)$ is given as

$$\widehat{\boldsymbol{\beta}}(x) = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{Y}.$$

Note that $\mathbf{X}$ is a Vandermonde matrix, and $\mathbf{W}$ is a diagonal matrix. One can therefore conclude that $\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}$ is regular as long as there are $p+1$ different local effective design points, i.e. the set $\left\{ X_i \colon K\left(\frac{X_i - x}{h_n}\right) \neq 0 \right\}$ contains at least $p+1$ elements. Since we will assume that $nh_n \underset{n\to\infty}{\longrightarrow} \infty$, this assumption is granted with probability tending to one.

The decomposition of $m(X_t)$ in equation (2.1) provides

$$\mathbf{Y} = \mathbf{m} + \boldsymbol{\sigma}\mathbf{U} = \mathbf{X}\,\boldsymbol{\beta}(x) + \frac{m^{(p+1)}(x)}{(p+1)!} h_n^{p+1}\,\mathbf{B}_n(x) + \boldsymbol{\sigma}\mathbf{U} + h_n^{p+1}\,\mathbf{e}_n(x),$$

with $\mathbf{m} := (m(X_1), \ldots, m(X_n))^\mathsf{T}$, $\boldsymbol{\sigma}\mathbf{U} := (\sigma(X_1)U_1, \ldots, \sigma(X_n)U_n)^\mathsf{T}$, and

$$\mathbf{B}_n(x) := \begin{pmatrix} \left(\frac{X_1-x}{h_n}\right)^{p+1} \\ \vdots \\ \left(\frac{X_n-x}{h_n}\right)^{p+1} \end{pmatrix}, \text{ and } \mathbf{e}_n(x) := \begin{pmatrix} \frac{m^{(p+1)}(\xi_1)-m^{(p+1)}(x)}{(p+1)!} \left(\frac{X_1-x}{h_n}\right)^{p+1} \\ \vdots \\ \frac{m^{(p+1)}(\xi_n)-m^{(p+1)}(x)}{(p+1)!} \left(\frac{X_n-x}{h_n}\right)^{p+1} \end{pmatrix}.$$

Therefore, we can decompose the local polynomial estimator in a bias term $\mathbf{B}_x$, a variance term $\mathbf{V}_x$, and an error term $\mathbf{e}_x$ as follows:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}(x) &= (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{Y} \\ &= \boldsymbol{\beta}(x) + h_n^{p+1}\frac{m^{(p+1)}(x)}{(p+1)!}(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{B}_n(x) + (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\boldsymbol{\sigma}\mathbf{U} + \mathbf{e}_x \\ &=: \boldsymbol{\beta}(x) + \mathbf{B}_x + \mathbf{V}_x + \mathbf{e}_x,\end{aligned}$$

where the error term $\mathbf{e}_x := h_n^{p+1}(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\ \mathbf{e}_n(x)$ is of asymptotically negligible order $o_P\left(h_n^{p+1}\right)$, see Corollaries 2.12 and 2.17.

2.2 Asymptotic normality under strong mixing conditions

We derive a joint asymptotic normality for $\widehat{\boldsymbol{\beta}}(x)$ and its derivatives:

$$\sqrt{nh_n}\left(\widehat{\boldsymbol{\beta}}(x) - \boldsymbol{\beta}(x) - h_n^{p+1}\mathbf{M}^{-1}\widetilde{\mathbf{B}}\,\frac{m^{(p+1)}(x)}{(p+1)!}\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(x)\sigma_U^2/f_X(x)\mathbf{M}^{-1}\boldsymbol{\Gamma}\mathbf{M}^{-1}\right),$$

where $f_X(x)$ denotes the marginal density of X in x, and the elements of the matrices $\mathbf{M}$ and $\widetilde{\mathbf{B}}$ are the moments of $K(\cdot)$

$$\mathbf{M} := \begin{pmatrix} \int K(u)du & \ldots & \int K(u)u^p du \\ \vdots & \ddots & \vdots \\ \int K(u)u^p du & \ldots & \int K(u)u^{2p}du \end{pmatrix}, \quad \widetilde{\mathbf{B}} := \begin{pmatrix} \int K(u)u^{p+1}du \\ \vdots \\ \int K(u)u^{2p}du \end{pmatrix},$$

and the matrix $\boldsymbol{\Gamma}$ denotes the moments of $K^2(\cdot)$

$$\boldsymbol{\Gamma} := \begin{pmatrix} \int K^2(u)du & \ldots & \int K^2(u)u^p du \\ \vdots & \ddots & \vdots \\ \int K^2(u)u^p du & \ldots & \int K^2(u)u^{2p}du \end{pmatrix}.$$

The estimator $\widehat{\boldsymbol{\beta}}(x)$ is thus an asymptotically unbiased estimator for $\boldsymbol{\beta}(x)$, and the order of the bias is $O(h_n^{p+1})$. Note that in the case of i.i.d. design points $X_t,\ t = 1, \ldots, n$ and a centered kernel, the local polynomial estimator is unbiased, e.g. $\mathbb{E}\left(\widehat{\boldsymbol{\beta}}(x)|X_1, \ldots, X_n\right) = \boldsymbol{\beta}(x)$. Further, the asymptotic variance is proportional to the variance σ_U^2 of the error process. We will make use of this fact later when we derive a more efficient estimator.

The classical results suppose the observations to be independent. In contrast to that, we want to allow for a certain kind of weak dependence. Heuristically, a time series is weak dependent if its values at widely separated times are asymptotically independent. In the last decades, mixing conditions have been the dominating way to formalize weak dependence. In the present thesis, we will make use of strong mixing conditions as the most general of all mixing conditions.

Our asymptotic results are based on the works of Masry (1996*a*,*b*) and Masry & Fan (1997) who derived asymptotic properties of local polynomial estimation for strongly mixing observations and established uniform convergence rates. In what follows, we demonstrate that these results remain valid in our regression setting. We first provide an overview of the strong mixing property, before we derive asymptotic expressions for the bias term $\mathbf{B}_x$ and for the variance term $\mathbf{V}_x$ under strong mixing conditions. We conclude this section by showing that these expressions imply the asymptotic normality of the local polynomial estimator.

2.2.1 Preliminaries

We start with a brief introduction of the strong mixing property. Strong mixing conditions and the associated central limit theorems have enjoyed broad appeal in the statistics and probability community. Various tools such as central limit theorems and moment inequalities can be carried over from the i.i.d. setting to mixing processes (Doukhan 1994, Bradley 2007). Loosely speaking, the strong mixing assumption is one of asymptotic independence: the statistical dependence between two random variables Z_{t_1} and Z_{t_2} as part of a random process $(Z_t)_{t\in\mathbb{Z}}$ goes to zero as their corresponding time gap $|t_2 - t_1|$ increases. To make this precise, we need to specify how we measure the dependence between Z_{t_1} and Z_{t_2}. In this thesis, we will use a very common choice introduced by Rosenblatt (1956):

Definition 2.1 (Strong mixing coefficients)**.** The strong mixing (or alpha-mixing) coefficients of a random process $(Z_t)_{t\in\mathbb{Z}}$ are defined as

$$\alpha(k) := \sup_{t\in\mathbb{Z}} \alpha\left(\sigma(Z_s, s \le t), \sigma(Z_s, s \ge t+k)\right), \text{ for all } k \in \mathbb{N}.$$

Here, $\sigma(Z_l, l \in T)$ denotes the σ-algebra generated by $(Z_l)_{l\in T}$ with $T \subset \mathbb{Z}$. The mixing coefficients $\alpha(\cdot,\cdot)$ of two sigma-algebras are defined as

$$\alpha(\mathcal{U}, \mathcal{V}) := \sup_{U\in\mathcal{U}, V\in\mathcal{V}} \left|\mathrm{P}(U \cap V) - \mathrm{P}(U)\mathrm{P}(V)\right|,$$

where $\mathcal{U}$ and $\mathcal{V}$ are sub-σ-algebras of the main σ-algebra $\mathcal{A}$ in a given probability space $(\Omega, \mathcal{A}, \mathrm{P})$. $\diamond$

We can now specify the definition of asymptotic independence:

Definition 2.2 (Strongly mixing process)**.** A random process $(Z_t)_{t\in\mathbb{Z}}$ is called strongly mixing (or $\alpha-$*mixing*), if $\alpha(k) \to 0$ as $k \to \infty$. $\diamond$

Remark 2.3. *Note that* $1/4 \geq \alpha(1) \geq \alpha(2) \geq \ldots \geq 0.$

The strong mixing assumption is the weakest of all well-known mixing conditions and covers a wide range of examples.

Example 2.4 (Strongly mixing processes)**.**

1. Obviously, i.i.d. - sequences are strongly mixing.

2. A basic example for a strongly mixing process whose mixing coefficients are geometrically decreasing is a stationary AR(q) sequence generated by continuous white noise:

$$Z_t = \alpha_1 Z_{t-1} + \ldots + \alpha_q Z_{t-q} + \varepsilon_t,$$

   with $\varepsilon_t \sim \mathcal{N}(0, 1)$ i.i.d, and the roots of the AR polynomial $f(x) := \alpha_1 x + \ldots + \alpha_q x^q$ outside the unit circle. More general, all stationary processes of ARMA type are geometrically strong mixing, provided that the innovations process $(\varepsilon_t)_{t\in\mathbb{Z}}$ possesses an absolutely continuous distribution with respect to the Lebesgue measure (Mokkadem 1988).[3]

3. Consider the following nonlinear process

$$Z_t = g_1(Z_{t-1}) + g_2(Z_{t-1})\varepsilon_t,$$

   with $(\varepsilon_t)_{t\in\mathbb{Z}}$ an i.i.d. random sequence with mean zero and finite variance, and a positive function $g_2(\cdot)$. If $g_1(\cdot)$ and $g_2(\cdot)$ are bounded or with growth at infinity slower than a linear function, then the process $(Z_t)_{t\in\mathbb{Z}}$ is stationary and strongly mixing with exponentially decreasing mixing coefficients (Masry & Tjøstheim 1995, 1997).

4. In general, all ergodic Markov Chains are strongly mixing (Doukhan 1994).

A broad variety of other examples such as gaussian random fields and diffusion processes can be found in Doukhan (1994) and Bradley (2007). As these authors also noted, nearly all processes that satisfy certain ergodicity and stability conditions are strongly mixing. $\diamond$

Throughout this thesis, we will make frequent use of the following inequality which combines the strong mixing property with the covariance of two variables:

[3] One of the most-cited counterexamples occurs if we substitute the underlying white noise process by a binomial one: Let $\varepsilon_t \sim \mathcal{B}_{1,1/2}$ i.i.d. Then the AR(1)-process $(Z_t)_{t\in\mathbb{Z}}$ with $0 \leq |\alpha_1| \leq 1/2$ does not satisfy the strong mixing property, as demonstrated by Andrews (1984).

Lemma 2.5 (Covariance inequality for strongly mixing random variables, Bradley (2007)). *Let X and Y be two random variables with $\|X\|_q < \infty$ and $\|Y\|_r < \infty$. Then*

$$|\mathrm{Cov}(X, Y)| \leq 4\, \alpha(\sigma(X), \sigma(Y))^{1/p} \|X\|_q \|Y\|_r,$$

for any $p, q, r > 1$ and $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1$.

Proof. See Bradley (2007, Corollary 10.16). □

Corollary 2.6 (Covariance inequality for strongly mixing random processes). *Let $(Z_t)_{t\in\mathbb{Z}}$ be a random process with $\|Z_t\|_q < \infty$, and $\|Z_{t+k}\|_r < \infty$. Then*

$$|\mathrm{Cov}(Z_t, Z_{t+k})| \leq 4\, \alpha(k)^{1/p} \|Z_t\|_q \|Z_{t+k}\|_r,$$

for any $p > 1, q \geq r > 1$ and $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1$.

Further, we will need the following result for the product of two mixing processes:

Lemma 2.7 (Product of independent strongly mixing processes, Bradley (2007)). *If $(X_t)_{t\in\mathbb{Z}}$ and $(Y_t)_{t\in\mathbb{Z}}$ are two independent sequences that are strongly mixing with mixing coefficients $\alpha_X(\cdot)$ and $\alpha_Y(\cdot)$, respectively, then the product process $(Z_t)_{t\in\mathbb{Z}}$ with $Z_t := X_t \cdot Y_t$ is also strongly mixing with mixing coefficients $\alpha_Z(\cdot)$ that satisfy $\alpha_Z(k) \leq \alpha_X(k) + \alpha_Y(k)$.*

Proof. The proof is a direct application of Bradley (2007, Lemma 6.4 (a)). □

We can now establish a joint convergence result for the local polynomial estimator and its derivatives. The following assumptions will facilitate our asymptotic analysis:

Assumption 1.

(i) The conditional mean function $m(\cdot)$ is $(p+1)$-times differentiable and the derivatives are continuous and bounded.

(ii) The function $\sigma^2(\cdot)$ and its inverse $1/\sigma^2(\cdot)$ are continuous and bounded.

(iii) The stationary process $(X_t)_{t\in\mathbb{Z}}$ is strongly mixing with mixing coefficients $\alpha_X(\cdot)$ that satisfy $\sum_{k=0}^{\infty} k^a \alpha_X(k)^{1-2/\mu} < \infty$ for some $\mu > 2$ and $a > (1 - 2/\mu)$. The density $f_X(\cdot)$ of $(X_t)_{t\in\mathbb{Z}}$ as well as the joint densities of $f_{X_{t_1},\ldots,X_{t_k}}(\cdot)$ of $(X_{t_1}, \ldots, X_{t_k})$ and the conditional densities $f_{(X_{t_1},\ldots,X_{t_k})|(X_{t_l},\ldots,X_{t_m})}(\cdot)$ of $(X_{t_1}, \ldots, X_{t_k})|(X_{t_l}, \ldots, X_{t_m})$ are continuous, bounded and bounded away from zero.

(iv) The stationary error process $(U_t)_{t\in\mathbb{Z}}$ has finite third moments, is independent of $(X_t)_{t\in\mathbb{Z}}$ and strongly mixing with mixing coefficients $\alpha_U(\cdot)$ that satisfy $\sum_{k=0}^{\infty} \alpha_U(k)^{1/3} < \infty$.

(v) The kernel $K(\cdot)$ is continuous, bounded, non-negative and strictly positive in zero, and has support on $[-1, 1]$.

Remark 2.8 (Some notes on the assumptions)**.** *Under the strong mixing Assumption 1 (iii) on the design points, the temporal dependence among $(X_t)_{t\in\mathbb{Z}}$ decreases sufficiently fast to be asymptotically ignorable. Furthermore, the so-called* whitening by windowing principle*, first introduced by Hart (1996), occurs: the kernel estimator uses only data points within the local window $x \pm h_n$. The dependence of the local series $\left\{X_{t(j)}, j = 1, \ldots, J\right\}$ in such a small window is much weaker than the one in the original dataset. This is due to the fact that the local time sequence $\{t(j), j = 1, \ldots, J\}$ is likely to be far apart. For a strongly mixing process the local series will hence be nearly independent. This enables us to carry over the asymptotic results for local polynomial estimation with independent observations.*

Assumption 1 (i) ensures a Taylor expansion of appropriate order, whereas under Assumption 1 (iv) the error dependence decreases sufficiently fast to derive the asymptotic variance of the local polynomial estimator. Assumption 1 (v) on the kernel function is a standard assumption in nonparametric statistics.

2.2.2 Asymptotic expression for the bias term

We first derive an asymptotic approximation for the bias term

$$\mathbf{B}_x = h_n^{p+1}\frac{m^{(p+1)}(x)}{(p+1)!}(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{B}_n \xrightarrow{\text{P}} h_n^{p+1}\frac{m^{(p+1)}(x)}{(p+1)!}\mathbf{M}^{-1}\widetilde{\mathbf{B}}, \tag{2.3}$$

by studying the convergence of $\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}$ and $\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{B}_n$. We derive pointwise, as well as uniform convergence results for these expressions. As a side effect, we show that the error term $\mathbf{e}_x$ is asymptotically negligible.

Lemma 2.9. *Under Assumption 1, the term $\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}$ converges in the mean-squared sense*

$$\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X} \xrightarrow{m.s.} f_X(x)\mathbf{M},$$

for $h_n \to 0, nh_n \to \infty$.

Proof. We prove the convergence element-wise with

$$b_j(x) := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - x}{h_n}\right)^j \frac{1}{h_n}K\left(\frac{X_i - x}{h_n}\right) = [\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}]_{k,l}, \text{ with } j = k + l - 2. \tag{2.4}$$

We show that $\mathbb{E}(b_j(x) - f_X(x)[\mathbf{M}]_{k,l})^2 \underset{n\to\infty}{\longrightarrow} 0$. It suffices to demonstrate the following convergences of the mean and the variance:

- $\mathbb{E}(b_j(x)) \underset{n\to\infty}{\longrightarrow} f_X(x)[\mathbf{M}]_{k,l}$

- $nh_n \mathrm{Var}(b_j(x)) \underset{n\to\infty}{\longrightarrow} f_X(x)\gamma_{2j}$ $(\Rightarrow \mathrm{Var}(b_j(x)) \underset{n\to\infty}{\longrightarrow} 0)$, with $\gamma_{2j} := \int K^2(u)u^{2j}du.$

Convergence of the mean. The stationarity of $(X_t)_{t\in\mathbb{Z}}$ provides

$$\begin{aligned}
&|\mathbb{E}(b_j(x)) - f_X(x)[\mathbf{M}]_{k,l}| \\
&= \left|\mathbb{E}\left(\left(\frac{X_1-x}{h_n}\right)^j \frac{1}{h_n}K\left(\frac{X_1-x}{h_n}\right)\right) - f_X(x)[\mathbf{M}]_{k,l}\right| \\
&= \left|\int \left(\frac{u-x}{h_n}\right)^j \frac{1}{h_n}K\left(\frac{u-x}{h_n}\right) f_X(u)du - f_X(x)\int y^j K(y)dy\right| \\
&= \left|\int (f_X(x+h_n y) - f_X(x))\, y^j K(y)dy\right| \\
&= o(1)\int |y|^j |K(y)|dy = o(1),
\end{aligned}$$

where the second last equality follows from dominated convergence, since the continuity of $f(\cdot)$ implies that $|f_X(x+h_ny) - f_X(x)| \to 0$ for $h_n \to 0$.

Convergence of the variance. Define

$$Z_{n,i} := \left(\frac{X_i-x}{h_n}\right)^j \frac{1}{h_n}K\left(\frac{X_i-x}{h_n}\right).$$

The stationarity of $(X_t)_{t\in\mathbb{Z}}$ again gives

$$\begin{aligned}
\mathrm{Var}(b_j(x)) &= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^n Z_{n,i}\right) \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(Z_{n,i}) + \frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{k=i+1}^n \mathrm{Cov}(Z_{n,i}, Z_{n,k}) \\
&= \frac{1}{n}\mathrm{Var}(Z_{n,1}) + \frac{2}{n}\sum_{i=2}^n\left(1 - \frac{i-1}{n}\right)\mathrm{Cov}(Z_{n,1}, Z_{n,i}) \\
&=: J_{n,1} + J_{n,2}.
\end{aligned} \tag{2.5}$$

For $J_{n,1}$, we have that

$$nJ_{n,1} = \mathrm{Var}(Z_{n,1}) = \mathbb{E}(Z_{n,1}^2) - (\mathbb{E}(Z_{n,1}))^2$$

The boundedness of the kernel function by Assumption 1 (v) provides

$$
\begin{aligned}
|\mathbb{E}(Z_{n,1})| &= \left| \int \left(\frac{u-x}{h_n}\right)^j \frac{1}{h_n} K\left(\frac{u-x}{h_n}\right) f_X(u) du \right| \\
&\leq \int_{x-h_n}^{x+hn} \left|\frac{u-x}{h_n}\right|^j \left|\frac{1}{h_n} K\left(\frac{u-x}{h_n}\right)\right| f_X(u) du \\
&\leq \int_{x-h_n}^{x+hn} \frac{C}{h_n} f_X(u) du \\
&\leq C,
\end{aligned} \tag{2.6}
$$

and again by dominated convergence we get

$$
\begin{aligned}
&\left| h_n \mathbb{E}(Z_{n,1}^2) - f_X(x) \int u^{2j} K^2(u) du \right| \\
&= \left| \int \left(\frac{v-x}{h_n}\right)^{2j} \frac{1}{h_n} K^2\left(\frac{v-x}{h_n}\right) f_X(v) dv - f_X(x) \int u^{2j} K^2(u) du \right| \\
&= \left| \int u^{2j} K^2(u) f_X(x + h_n u) du - f_X(x) \int u^{2j} K^2(u) du \right| \\
&\leq \int \left| u^{2j} K^2(u) (f_X(x + h_n u) - f_X(x)) \right| du \\
&= \ o(1) \int \left| u^{2j} K^2(u) \right| du \\
&= o(1).
\end{aligned}
$$

Hence,

$$
n h_n J_{n,1} = f_X(x) \gamma_{2j} + o(1). \tag{2.7}
$$

We decompose $J_{n,2}$ into a finite sum and a tail sum as follows:

$$
\begin{aligned}
J_{n,2} &= \frac{2}{n} \sum_{2 \leq i \leq \pi_n} \left(1 - \frac{i-1}{n}\right) \operatorname{Cov}(Z_{n,1}, Z_{n,i}) + \frac{2}{n} \sum_{i \geq \pi_n}^{n} \left(1 - \frac{i-1}{n}\right) \operatorname{Cov}(Z_{n,1}, Z_{n,i}) \\
&=: J_{n,21} + J_{n,22},
\end{aligned} \tag{2.8}
$$

with π_n such that $\pi_n \underset{n\to\infty}{\longrightarrow} \infty$, and $h_n \pi_n \underset{n\to\infty}{\longrightarrow} 0$. The first sum is small by the choice of π_n. Further, the strong mixing property of $(X_t)_{t\in\mathbb{Z}}$ yields that the tail sum is asymptotically negligible, since the time lag between Z_1 and all Z_i is sufficiently big.

For $J_{n,21}$, we get

$$
n|J_{n,21}| \leq 2 \sum_{2 \leq i \leq \pi_n} |\operatorname{Cov}(Z_{n,1}, Z_{n,i})|.
$$

Since the joint densities are bounded and the kernel is bounded with compact support by Assumption 1, we have

$$\begin{aligned}|\mathbb{E}(Z_{n,1}Z_{n,i})| &= \left|\int \frac{1}{h_n^2}\left(\frac{u-x}{h_n}\right)^j \left(\frac{v-x}{h_n}\right)^j K\left(\frac{u-x}{h_n}\right) K\left(\frac{v-x}{h_n}\right) f_X(u,v)dudv\right| \\ &\le C\left|\int w^j z^j K(w)\,K(z)\,dwdz\right| \le C.\end{aligned}$$

Since $|\mathbb{E}(Z_{n,1})| \le C$ from equation (2.6), we get $|\mathrm{Cov}(Z_{n,1}, Z_{n,i})| \le C$. Hence by the choice of π_n

$$nh_n J_{n,21} = O(h_n \pi_n) = o(1). \tag{2.9}$$

For $J_{n,22}$, we have $n|J_{n,22}| \le 2\sum_{i=\pi_n}^{n} |\mathrm{Cov}(Z_{n,1}, Z_{n,i})|$. Since X is an α- mixing process we get from the covariance inequality for mixing processes as stated in Lemma 2.5

$$|\mathrm{Cov}(Z_{n,1}, Z_{n,i})| \le 4\,\alpha_X(i-1)^{1-2/\mu}\|Z_{n,1}\|_\mu^2.$$

It holds

$$|\mathbb{E}(Z_{n,1}^\mu)| = \left|\int \left(\frac{u-x}{h_n}\right)^{\mu j} \frac{1}{h_n^\mu} K^\mu\left(\frac{u-x}{h_n}\right) f_X(u)du\right| = O\left(\frac{1}{h_n^{\mu-1}}\right),$$

uniformly in x, again by the boundedness of the density $f_X(\cdot)$ and the boundedness and compact support property of the Kernel. Hence $\|Z_{n,1}\|_\mu^2 = O(h_n^{2/\mu-2})$, and thus

$$nh_n|J_{n,22}| \le \frac{C}{h_n^{(1-2/\mu)}} \sum_{i\ge\pi_n}^{n} \alpha_X(i-1)^{1-2/\mu} \le \frac{C}{h_n^{(1-2/\mu)}\pi_n^a} \sum_{i\ge\pi_n}^{n} i^a \alpha_X(i-1)^{1-2/\mu}.$$

Set $\pi_n = h_n^{\frac{2/\mu-1}{a}}$, and note that $h_n^{(1-2/\mu)}\pi_n^a = 1$, and as well $h_n\pi_n \underset{n\to\infty}{\longrightarrow} 0$ as required.[4] By Assumption 1 (iii) the partial sum $\sum_{i=\pi_n}^{n} i^a \alpha_X(i-1)^{1-2/\mu}$ converges to zero, and hence

$$nh_n J_{n,22} = o(1). \tag{2.10}$$

Equations (2.5) - (2.10) yield the assertion. □

The convergence of the inverse $(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}$ is a direct implication of Lemma 2.9:

Corollary 2.10. *Let Assumption 1 hold. If $f_X(x) > 0$, we have*

$$\mathrm{P}\left(\text{"}\mathbf{X}^\top\mathbf{W}\mathbf{X} \textit{ not inverible or } \left\|(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1} - \mathbf{M}^{-1}f_X(x)^{-1}\right\| > \delta_n\text{"}\right) \underset{n\to\infty}{\longrightarrow} 0,$$

for an arbitrary null sequence $(\delta_n)_{n\ge1}$. Further, $\mathbf{M}$ *is a symmetric and positive definite matrix.*

[4]It holds $h_n\pi_n = h_n^{1+\frac{2/u-1}{a}} = h_n^{\frac{a+2/\mu-1}{a}}$, and since $a > (1-2/\mu)$ by Assumption 1 (iii) we get $h_n\pi_n = h_n^\rho$, with $\rho > 0$.

Proof. Obviously, $\mathbf{M}$ is symmetric. We now demonstrate that $\mathbf{M}$ is positive definite (and thus invertible). For $0 \neq c \in \mathbb{R}^{p+1}$ arbitrary we get

$$c^\mathsf{T}\mathbf{M}c = \int_{-1}^{1} K(u) \left(\sum_{i=0}^{p} c_i u^i \right)^2 du > 0,$$

since by Assumption 1 (v) the kernel function $K(\cdot)$ is continuous, non-negative and $K(0) > 0$, which implies $K(u) > 0$ for all $|u| < \varepsilon$, for some $\varepsilon > 0$. $\mathbf{M}$ is hence invertible and now the last Lemma directly provides

$$\mathrm{P}(\text{" } \mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X} \text{ is positive definite"}) \underset{n\to\infty}{\longrightarrow} 1.$$

Thus we now suppose that $\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}$ is positive definite for all $n \geq N_0$. It remains to demonstrate the convergence of $(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}$. The last Lemma provides $\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X} \xrightarrow{\mathrm{P}} \mathbf{M}f_X(x)$, since mean square convergence implies convergence in probability. From the continuous mapping theorem (see for example Shao (2003)) we get that for a set of random matrices $(\mathbf{Z}_t)_{t\in\mathbb{Z}}$ with $\mathbf{Z}_t \xrightarrow{P} \mathbf{A}$ and a continuous function $g\colon \mathbb{R}^{m\times m} \to \mathbb{R}^{m\times m}$, the sequence $(g(\mathbf{Z}_t))_{t\in\mathbb{Z}}$ satisfies $g(\mathbf{Z}_t) \xrightarrow{P} g(\mathbf{A})$. Since the matrix inverse is a continuous function over the space of nonsingular matrices, we hence conclude

$$(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1} \xrightarrow{\mathrm{P}} \mathbf{M}^{-1}f_X(x)^{-1},$$

which yields the assertion. □

Remark 2.11 (Regularity of the matrix $\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}$)**.** *Regarding the previous results, from now on we will assume that* $\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}$ *is positive definite for all* $n \in \mathbb{N}$.

Similarly to Lemma 2.9, we calculate the stochastic convergence

$$\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{B}_n \xrightarrow{\mathrm{P}} f_X(x)\widetilde{\mathbf{B}}.$$

Thus the asymptotic convergence of the bias as formulated in equation (2.3) holds true. Furthermore, we conclude that the error term $\mathbf{e}_x$ is asymptotically negligible:

Corollary 2.12. *Under Assumption 1, it holds*

$$\mathbf{e}_x = o_P\left(h_n^{p+1}\right).$$

Proof. Since the $(p+1)$th derivative of $m(\cdot)$ is bounded by Assumption 1 (i) and the kernel function assures that $|\xi_i - x| \leq h_n$, we have $|m^{(p+1)}(\xi_i) - m^{(p+1)}(x)| \leq Ch_n$. Hence applying

Markov's inequality on the remaining sum term, we obtain

$$\left[\mathbf{X}^{\mathsf{T}}\mathbf{W}\,\mathbf{e}_n(x)\right]_j \leq C h_n \frac{1}{nh_n}\sum_{i=1}^{n}\left|\left(\frac{X_i - x}{h_n}\right)^{p+j} K\left(\frac{X_i - x}{h_n}\right)\right| = O_P(h_n).$$

Now since $(\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X})^{-1} = O_P(1)$ by Corollary 2.10, we get

$$\mathbf{e}_x = h_n^{p+1}(\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{W}\,\mathbf{e}_n(x) = h_n^{p+1}\,O_P(1)\,O_P(h_n) = O_P(h_n^{p+2}) = o_P(h_n^{p+1}).$$

□

2.2.3 Uniform convergence rate for the bias term

In the following chapters, our calculations will be facilitated by the usage of convergence rates for the matrix $\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X}$ and the bias term $\mathbf{B}_x$ that hold uniformly over compact intervals. In order to derive such uniform convergence results, we make use of the following auxiliary Lemma contributed by Masry:

Lemma 2.13 (Masry (1996*a*))**.** *Let $D \subset \mathbb{R}$ be a compact interval, and let Assumption 1 hold. Assume the bandwidth $h_n \underset{n\to\infty}{\longrightarrow} 0$ such that $nh_n/\log(n) \underset{n\to\infty}{\longrightarrow} \infty$. If the strong mixing coefficients $\alpha_X(\cdot)$ satisfy*

$$\sum_{n=1}^{\infty} \frac{n^{5/4}}{h_n^{9/4}(\log(n))^{1/4}}\,\alpha_X\left(\left(\frac{nh_n}{\log(n)}\right)^{1/2}\right) < \infty,$$

then for each $0 \leq j \leq 2p$ we have

$$\sup_{x\in D}|b_j(x) - \mathbb{E}(b_j(x))| = O\left(\left(\frac{\log n}{nh_n}\right)^{1/2}\right) \quad \textit{almost surely,}$$

with $b_j(x)$ as defined in equation (2.4).

Proof. We adapt the proof of Masry (1996*a*, proof of Theorem 2). We only give a short outline of the proof here, mainly demonstrating how to derive the basic steps under our assumptions.

Since D is compact, it can be covered by a finite number $L = L(n)$ of intervals $I_k = I_{n,k}$ with length l_n and centers $x_k = x_{n,k}$. Clearly, $l_n = C/L(n)$. We write

$$
\begin{aligned}
&\sup_{x\in D} |b_j(x) - \mathbb{E}(b_j(x))| \\
&= \max_{1\le k\le L(n)} \sup_{x\in D\cap I_k} |b_j(x) - \mathbb{E}(b_j(x))| \\
&\le \max_{1\le k\le L(n)} \sup_{x\in D\cap I_k} |b_j(x) - b_j(x_k)| + \max_{1\le k\le L(n)} |b_j(x_k) - \mathbb{E}(b_j(x_k))| \\
&\quad + \max_{1\le k\le L(n)} \sup_{x\in D\cap I_k} |\mathbb{E}(b_j(x_k)) \ - \mathbb{E}(b_j(x)) \ | \\
&=: \ Q_{1,n} + Q_{2,n} + Q_{3,n}.
\end{aligned}
$$

Since the kernel is Lipschitz, bounded and has compact support, it holds

$$
Q_{1,n} \le \max_{1\le k\le L(n)} \sup_{x\in D\cap I_k} \frac{C}{h_n}|x - x_k| \le \frac{C\ l_n}{h_n} = O\left(\left(\frac{\log n}{nh_n}\right)^{1/2}\right),
$$

by an adequate choice of l_n. Similarly, $Q_{3,n} = O\left(\left(\frac{\log n}{nh_n}\right)^{1/2}\right)$.

The main task is to show that $Q_{2,n} = O\left(\left(\frac{\log n}{nh_n}\right)^{1/2}\right)$ almost surely. In later chapters, we will make use of a Bernstein-type inequality for strongly mixing processes to derive asymptotic orders of similar terms. However, since the Bernstein-inequality assumes the strong mixing coefficients to be exponentially decreasing, we will prove the assertion directly in order to be less restrictive.

Write

$$
W_n = b_j(x) - \mathbb{E}(b_j(x)) = \frac{1}{n}\sum_{i=1}^{n} Z_{n,i},
$$

where

$$
Z_{n,i} := \frac{1}{h_n}\left(\left(\frac{X_i - x}{h_n}\right)^j K\left(\frac{X_i - x}{h_n}\right) - \mathbb{E}\left(\left(\frac{X_i - x}{h_n}\right)^j K\left(\frac{X_i - x}{h_n}\right)\right)\right).
$$

We partition the set $\{0, 1, \ldots, n\}$ into $2q = 2q(n)$ blocks of size $r = r(n) = \left\lfloor\sqrt{\frac{nh_n}{\log n}}\right\rfloor$; $n = rq + v$, with $0 \le v = v(n) \le r$, and write

$$
V_n(j) = \frac{1}{n}\sum_{i=(j-1)r+1}^{jr} Z_{n,i}, \quad j = 1, \ldots, 2q,
$$

and $W_n = W_n' + W_n'' + W_n'''$, with the sums of odd-numbered blocks W_n' and of even-numbered blocks $W_n^{''}$ defined as

$$W_n^{'} := \sum_{j=1}^{q} V_n(2j-1), \;\; W_n^{''} := \sum_{j=1}^{q} V_n(2j), \;\; W_n^{'''} := \sum_{i=2qr+1}^{n} Z_{n,i}.$$

The remainder term $W_n^{'''}$ is asymptotically negligible, since it consists of at most $r(n)$ terms, whereas each term of $W_n^{'}$ and $W_n^{''}$ contains $q(n)r(n)$ elements, with $q(n) \to \infty$. Further, stationarity provides

$$\begin{aligned} \mathrm{P}\left(Q_{2,n} > \delta\right) &\leq \mathrm{P}\left(\max_{1\leq k\leq L(n)} |W_n^{'}(x_k)| > \delta/2\right) + \mathrm{P}\left(\max_{1\leq k\leq L(n)} |W_n^{''}(x_k)| > \delta/2\right) \\ &\leq 2\, L(n) \sup_{x\in D} \mathrm{P}\left(|W_n^{'}(x)| > \delta/2\right). \end{aligned}$$

To bound the expression $\mathrm{P}\left(|W_n^{'}(x)| > \delta/2\right)$, the strong approximation theorem (Bradley 1983, p. 69-81) is used to approximate the series $(V_n(2j-1))_{j=1,\dots,q}$ by a series of independent variables $(V_n^*(2j-1))_{j=1,\dots,q}$ with the same distribution as $V_n(2j-1)$, enlarging the probability space if necessary. It holds

$$\begin{aligned} &\mathrm{P}\left(|W_n^{'}(x)| > \delta/2\right) \\ &\leq \mathrm{P}\left(\left|\sum_{j=1}^{q} V_n^*(2j-1)\right| > \delta/4\right) + \mathrm{P}\left(\left|\sum_{j=1}^{q} V_n(2j-1) - V_n^*(2j-1)\right| > \delta/4\right). \end{aligned}$$

We show that

$$\sup_{x\in D} \sum_{j=1}^{q} \mathbb{E}\left(V_n^*(2j-1)\right)^2 = O\left(\frac{1}{nh_n}\right). \tag{2.11}$$

The proof is then completed in the manner of Masry (1996*a*, proof of Theorem 2): the first sum is bounded by equation (2.11) whereas the second term is sufficiently small thanks to the assumed decay rate for $\alpha(\cdot)$.

Since $(V_n^*(2j-1))_{j=1,\dots,q}$ have the same distribution as $V_n(2j-1)$, we have

$$\begin{aligned} &\sum_{j=1}^{q} \mathbb{E}\left(V_n^*(2j-1)\right)^2 = \sum_{j=1}^{q} \mathbb{E}\left(V_n(2j-1)\right)^2 = \frac{1}{n^2}\sum_{j=1}^{q} \mathbb{E}\left(\sum_{i=(2j-1)r+1}^{2rj} Z_{n,i}\right)^2 \\ &\leq \frac{1}{n} \operatorname{Var}\left(Z_{n,i}\right) + \frac{1}{n^2} \sum_{i\neq j} \operatorname{Cov}\left(Z_{n,i}, Z_{n,j}\right). \end{aligned}$$

Now by Lemma 2.9 we have that

$$\begin{aligned}
&\sup_{x\in D} \operatorname{Var}\left(Z_{n,1}\right) \\
&\leq \frac{1}{h_n}\int_u \left|u^{2j}K^2(u)\right| \sup_{x\in D}|f_X(x+h_nu)-f_X(x)|du + \left(\sup_{x\in D}\int_{x-h_n}^{x+h_n}\frac{C}{h_n}f_X(u)du\right)^2 \\
&= O\left(\frac{1}{h_n}\right),
\end{aligned}$$

since $f_X(\cdot)$ is continuous and bounded by Assumption 1 (iii). Similarly,

$$\sup_{x\in D}\frac{1}{n^2}\sum_{i\neq j}\operatorname{Cov}\left(Z_{n,i},Z_{n,j}\right) = o\left(\frac{1}{nh_n}\right),$$

which yields the assertion. □

We can now state a convergence result for $\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}$ that holds uniformly over compact intervals.

Lemma 2.14 (Uniform convergence of $\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}$ over compact intervals)**.** *Let $D\subset\mathbb{R}$ be any compact interval, and let Assumption 1 hold. Assume further that $h_n = O\left(n^{-\gamma}\right)$ for some $0<\gamma<1$, and that the strong mixing coefficients of $(X_t)_{t\in\mathbb{Z}}$ satisfy $\alpha_X(k)=O\left(k^{-c}\right)$, for some $c>\frac{10}{1-\gamma}$. Then it holds*

$$\sup_{x\in D}|[\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}]_{k,l}-f_X(x)[\mathbf{M}]_{k,l}| = o(1) \ \ \textit{almost surely.}$$

Proof. As in the proof of Lemma 2.9, we have with $b_j(x)=[\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}]_{k,l}$ as defined before

$$|\mathbb{E}(b_j(x))-f_X(x)[\mathbf{M}]_{k,l}| \leq \int |(f_X(x+h_ny)-f_X(x))|\left|y^jK(y)\right|dy.$$

Since the density of the design process $f_X(\cdot)$ is bounded and continuous, it holds $|(f_X(x+h_ny)-f_X(x))| \underset{n\to\infty}{\longrightarrow} 0$ uniformly in $x\in D$. Hence

$$\sup_{x\in D}|\mathbb{E}(b_j(x))-f_X(x)[\mathbf{M}]_{k,l}| = o(1) \tag{2.12}$$

by dominated convergence.

We want to apply Lemma 2.13 to show that

$$\sup_{x\in D}|b_j(x)-\mathbb{E}(b_j(x))| = o\left(1\right) \ \ \text{almost surely.} \tag{2.13}$$

Note that $nh_n/\log(n) = O\left(n^{1-\gamma}\right)/\log(n) \to \infty$. Further, the mixing coefficients decay sufficiently fast:

$$\Psi(n) := \frac{n^{5/4}}{h_n^{9/4}(\log(n))^{1/4}}\,\alpha_X\left(\left(\frac{nh_n}{\log(n)}\right)^{1/2}\right) = \frac{n^{5/4}}{h_n^{9/4}(\log(n))^{1/4}} O\left(\left(\frac{nh_n}{\log(n)}\right)^{-c/2}\right),$$

and since $h_n = O\left(n^{-\gamma}\right)$ and $c > 10/(1-\gamma)$, this implies[5]

$$\begin{aligned}\Psi(n) &= O\left(\frac{n^{5/4}}{n^{-9\gamma/4}(\log(n))^{1/4}}\left(\frac{n^{1-\gamma}}{\log(n)}\right)^{-\frac{5}{1-\gamma}}\right) = O\left(n^{-5+5/4+9\gamma/4}\log(n)^{\left(\frac{5}{1-\gamma}-1/4\right)}\right)\\ &= o\left(n^{-3/2}\right).\end{aligned}$$

Thus $\sum_{n=1}^{\infty}\Psi(n) < \infty$. Lemma 2.13 now gives $\sup_{x\in D}|b_j(x) - \mathbb{E}(b_j(x))| = O\left(\sqrt{\frac{\log n}{nh_n}}\right)$ almost surely.

Since

$$|b_j(x) - f_X(x)[\mathbf{M}]_{k,l}| \leq |b_j(x) - \mathbb{E}(b_j(x))| + |\mathbb{E}(b_j(x)) - f_X(x)[\mathbf{M}]_{k,l}|,$$

the assertion follows from equations (2.12) and (2.13). □

The uniform convergence of the inverse, and as well the uniform rates of the bias and the error terms follow directly from the last Lemma.

Corollary 2.15 (Uniform convergence of $(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}$ over compact intervals)**.** *Under the assumptions of Lemma 2.14 we have, uniformly in $x \in D$,*

$$(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1} \underset{n\to\infty}{\longrightarrow} \mathbf{M}^{-1}f_X(x)^{-1} \text{ almost surely.}$$

Proof. The proof follows directly from Lemma 2.14 and the continuous mapping theorem in analogy to the proof of Corollary 2.10 . □

Corollary 2.16 (Uniform convergence rate for the bias term)**.** *Under the assumptions of Lemma 2.14, the following uniform bound for the bias term holds true:*

$$\sup_{x\in D}\mathbf{B}_x = O_P\left(h_0^{p+1}\right).$$

Proof. Similar calculations as for $\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}$ provide that $\sup_{x\in D}\left|\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{B}_n - f_X(x)\mathbf{B}\right| = o_P(1)$, which yields the assertion together with the last Corollary. □

[5]It was used that $\log(n) = o\left(n^a\right)$, for all $a > 0$ arbitrary.

Corollary 2.17 (Asymptotic negligibility of the error term uniformly over compact intervals)**.** *Under the assumptions of Lemma 2.14, the error term is asymptotically negligible, and this result holds uniformly over compact intervals:*

$$\sup_{x\in D} \mathbf{e}_x = o_P\left(h_0^{p+1}\right).$$

2.2.4 Asymptotic expression for the variance term

In order to derive the joint normality for $\widehat{\boldsymbol{\beta}}(x)$, we aim to establish the following convergence result for the variance term:

$$\operatorname{Cov}\left(\sqrt{nh_n}\mathbf{X}^\mathsf{T}\mathbf{W}\boldsymbol{\sigma}\mathbf{U}\right) \underset{n\to\infty}{\longrightarrow} \sigma_U^2\sigma^2(x)f_X(x)\boldsymbol{\Gamma}.$$

We show the convergence element-wise:

$$\operatorname{Cov}\left(\sqrt{nh_n}\, t_k, \sqrt{nh_n}\, t_l\right) \underset{n\to\infty}{\longrightarrow} \sigma_U^2\sigma^2(x)f_X(x)\gamma_{k+l}, \tag{2.14}$$

with t_j the $(j+1)$th element of $\mathbf{X}^\mathsf{T}\mathbf{W}\boldsymbol{\sigma}\mathbf{U}$

$$t_j := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i-x}{h_n}\right)^j \frac{1}{h_n}K\left(\frac{X_i-x}{h_n}\right)\sigma(X_i)U_i,\ j=0,\dots,p,$$

and $\gamma_{k+l} := [\Gamma]_{k+1,l+1} = \int u^{k+l}K^2(u)du$.

In order to prove equation (2.14), we adopt an auxiliary lemma from Masry (1996*b*). For this purpose, denote

$$Z_{n,i} := U_i\sigma(X_i)A_{h_n}(X_i-x), \tag{2.15}$$

where A_{h_n} is defined as

$$A_{h_n}(u) := \sum_{j=0}^{p} a_j\left(\frac{u}{h_n}\right)^j \frac{1}{h_n}K\left(\frac{u}{h_n}\right),$$

and Q_n the mean of the $Z_{n,i}$

$$Q_n := \frac{1}{n}\sum_{i=1}^{n} Z_{n,i}. \tag{2.16}$$

Further, let

$$A(z) := \sum_{j=0}^{p} a_j z^j K(z),$$

Then we get for the asymptotic variance of Q_n:

Lemma 2.18. *Under Assumption 1, it holds*

$$nh_n \mathrm{Var}(Q_n) \underset{n\to\infty}{\longrightarrow} \sigma_U^2 \sigma^2(x) f_X(x) \int A^2(z)dz,$$

for $h_n \to 0$, $nh_n \to \infty$.

Proof. We first establish the asymptotic variance of a single $Z_{n,i}$, and then that of the whole term.

Asymptotic variance of $Z_{n,i}$. We have

$$\begin{aligned}
\mathrm{Var}(Z_{n,i}) &= \mathbb{E}(Z_{n,i}^2) \\
&= \mathbb{E}(U_i^2 \sigma^2(X_i) A_{h_n}^2(X_i - x)) \\
&= \sigma_U^2 \mathbb{E}(\sigma^2(X_i) A_{h_n}^2(X_i - x)) \\
&= \sigma_U^2 \mathbb{E}\left(\frac{1}{h_n^2} \sum_{j,k=0}^{p} \sigma^2(X_i) a_j a_k \left(\frac{X_i - x}{h_n}\right)^{j+k} K^2\left(\frac{X_i - x}{h_n}\right)\right) \\
&= \sigma_U^2 \frac{1}{h_n} \sum_{j,k=0}^{p} a_j a_k \int \frac{1}{h_n} \left(\frac{u - x}{h_n}\right)^{j+k} K^2\left(\frac{u - x}{h_n}\right) \sigma^2(u) f_X(u) du.
\end{aligned}$$

Further, note that for $l \in \{0, \ldots, 2p\}$ arbitrary it holds

$$\begin{aligned}
&\left| \frac{1}{h_n} \int \left(\frac{u - x}{h_n}\right)^{l} K^2\left(\frac{u - x}{h_n}\right) \sigma^2(u) f_X(u) du - \sigma^2(x) f_X(x) \int z^l K^2(z) dz \right| \\
&= \left| \int \left(\sigma^2(x + h_n z) f_X(x + h_n z) - \sigma^2(x) f_X(x)\right) z^l K^2(z) dz \right| \\
&\leq \int \left|\sigma^2(x + h_n z) f_X(x + h_n z) - \sigma^2(x) f_X(x)\right| \left|z^l K^2(z)\right| dz.
\end{aligned}$$

Since the density function of the design process $f_X(\cdot)$ as well as the variance function $\sigma^2(\cdot)$ are continuous and bounded by Assumptions 1 (ii) and 1 (iii) we get further

$$\begin{aligned}
&\left|\sigma^2(x + h_n z) f_X(x + h_n z) - \sigma^2(x) f_X(x)\right| \\
&\leq \left|\sigma^2(x + h_n z) - \sigma^2(x)\right| \left|f_X(x + h_n z) - f_X(x)\right| + \left|\sigma^2(x)\right| \left|f_X(x + h_n z) - f_X(x)\right| \\
&\quad + \left|f_X(x)\right| \left|\sigma^2(x + h_n z) - \sigma^2(x)\right| \\
&= o(1),
\end{aligned}$$

and therefore by dominated convergence

$$\begin{aligned}
h_n \mathrm{Var}(Z_{n,i}) &= \sigma_U^2 \sum_{j,k=0}^{p} a_j a_k \int \frac{1}{h_n} \left(\frac{u-x}{h_n}\right)^{j+k} K^2\left(\frac{u-x}{h_n}\right) \sigma^2(u) f_X(u) du \\
&= \sigma_U^2 \sum_{j,k=0}^{p} a_j a_k \left(\sigma^2(x) f_X(x) \int z^{j+k} K^2(z)\, du\right)(1+o(1)) \\
&= \left(\sigma_U^2 \sigma^2(x) f_X(x) \int A^2(z) dz\right)(1+o(1)).
\end{aligned}$$

Hence

$$h_n \ \mathrm{Var}(Z_{n,i}) \underset{n\to\infty}{\longrightarrow} \sigma_U^2 \sigma^2(x) f_X(x) \int A^2(z) dz \tag{2.17}$$

Asymptotic variance of Q_n. By stationarity of $(X_t)_{t\in\mathbb{Z}}$ and $(U_t)_{t\in\mathbb{Z}}$ we have

$$\mathrm{Var}(Q_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Z_{n,i}\right) = \frac{1}{n}\mathrm{Var}(Z_{n,1}) + \frac{2}{n}\sum_{k=2}^{n}\left(1 - \frac{k-1}{n}\right)\mathrm{Cov}(Z_{n,1}, Z_{n,k}).$$

We show that

$$\sum_{k=2}^{n} |\mathrm{Cov}(Z_{n,1}, Z_{n,k})| = O(1).$$

Since $(X_t)_{t\in\mathbb{Z}}$ and $(U_t)_{t\in\mathbb{Z}}$ are independent we have

$$\mathrm{Cov}(Z_{n,1}, Z_{n,k}) = \mathbb{E}(Z_{n,1} Z_{n,k}) = \mathbb{E}(U_1 U_k)\mathbb{E}(\sigma(X_1)\sigma(X_k) A_{h_n}(X_1 - x) A_{h_n}(X_k - x)).$$

We make use of the mixing assumptions on $(U_t)_{t\in\mathbb{Z}}$. Since $(U_t)_{t\in\mathbb{Z}}$ is alpha-mixing, the covariance inequality for strongly mixing processes (Lemma 2.5) provides

$$|\mathbb{E}(U_1 U_k)| \le 4\ \alpha_U(k-1)^{1/3} \sup_s \|U_s\|_3^2 \le C\ \alpha_U(k-1)^{1/3},$$

since the third moments of $(U_t)_{t\in\mathbb{Z}}$ exist.

For the second product term we get

$$
\begin{aligned}
&|\mathbb{E}(\sigma(X_1)\sigma(X_k)A_{h_n}(X_1-x)A_{h_n}(X_k-x))| \\
&= \left|\mathbb{E}\left(\frac{1}{h_n^2}\sum_{j,l=0}^{p} a_j a_l \sigma(X_1)\sigma(X_k)\left(\frac{X_1-x}{h_n}\right)^j K\left(\frac{X_1-x}{h_n}\right)\left(\frac{X_k-x}{h_n}\right)^l K\left(\frac{X_k-x}{h_n}\right)\right)\right| \\
&\leq C\left(\int \left|\sigma(x+h_nu)\sigma(x+h_nv)u^j v^l K(u)K(v)\right| f_{X_1,X_k}(x+h_nu, x+h_nv)dudv\right) \\
&\leq C,
\end{aligned}
$$

since by Assumption 1 the joint density of X_1 and X_k is bounded, the kernel function $K(\cdot)$ is bounded with compact support and the variance function $\sigma^2(\cdot)$ is bounded. Since the mixing coefficients $\alpha_U(\cdot)^{1/3}$ of $(U_t)_{t\in\mathbb{Z}}$ are summable, we thus obtain

$$\sum_{k=2}^{n}|\mathrm{Cov}(Z_{n,1}, Z_{n,k})| \leq C\sum_{k=2}^{n-1}\alpha_U(k-1)^{1/3} = O(1),$$

and therefore

$$nh_n\mathrm{Var}(Q_n) = h_n\mathrm{Var}(Z_{n,1}) + O(h_n),$$

which yields the assertion together with equation (2.17). □

The asymptotic behavior of $\mathbf{V}_x$ is now a direct consequence of Lemma 2.18:

Lemma 2.19. *Under Assumption 1, it holds*

$$\mathrm{Cov}\left(\sqrt{nh_n}\, t_k, \sqrt{nh_n}\, t_l\right) \underset{n\to\infty}{\longrightarrow} \sigma_U^2\sigma^2(x)f_X(x)\gamma_{k+l},$$

for $h_n \to 0, nh_n \to \infty$.

Proof. For $a_m := \delta_k(m), m = 0, \ldots, p$ with the corresponding $Q_n^k := t_k$, Lemma 2.18 yields

$$\mathrm{Var}\left(\sqrt{nh_n}t_k\right) = nh_n\mathrm{Var}\left(Q_n^k\right) \underset{n\to\infty}{\longrightarrow} \sigma_U^2\sigma^2(x)f_X(x)\int u^{2k}K^2(u)\,du.$$

On the other hand, with $\widetilde{s}_m := \delta_k(m)+\delta_l(m), m = 1, \ldots, p$ and the corresponding $Q_n^{k,l} := t_k+t_l$ we get

$$nh_n\mathrm{Var}\left(Q_n^{k,l}\right) \underset{n\to\infty}{\longrightarrow} \sigma_U^2\sigma^2(x)f_X(x)\left(\int\left(u^{2k}+2u^{k+l}+u^{2l}\right)K^2(u)\,du\right)$$

And since $nh_n\mathrm{Var}\left(Q_n^{k,l}\right) = \mathrm{Var}\left(\sqrt{nh_n}t_k\right) + \mathrm{Var}\left(\sqrt{nh_n}t_l\right) + 2\mathrm{Cov}\left(\sqrt{nh_n}t_k, \sqrt{nh_n}t_l\right)$, we have

$$\operatorname{Cov}\left(\sqrt{nh_n}t_k, \sqrt{nh_n}t_l\right) \underset{n\to\infty}{\longrightarrow} \sigma_U^2\sigma^2(x)f_X(x)\int u^{k+l}K^2(u)\,du.$$

$\square$

2.2.5 Asymptotic normality

With the help of the asymptotic expressions for the bias term and the covariance matrix we can now develop a joint asymptotic normality for $\widehat{\boldsymbol{\beta}}(x)$. We make use of the following auxiliary convergence result:

Lemma 2.20 (Masry (1996*b*))**.** *Suppose Assumption 1 to hold. Assume further that* $h_n \underset{n\to\infty}{\longrightarrow} 0, nh_n \underset{n\to\infty}{\longrightarrow} \infty$, *and that the strong mixing coefficients of* $(X_t)_{t\in\mathbb{Z}}$ *satisfy* $\sqrt{n/h_n}\,\alpha_X\left(\sqrt{nh_n}\right) \underset{n\to\infty}{\longrightarrow} 0$. *Then for* $n \underset{n\to\infty}{\longrightarrow} \infty$ *we have*

$$\sqrt{nh_n}Q_n \xrightarrow{D} \mathcal{N}\left(0, \sigma_U^2\sigma^2(x)f_X(x)\int A^2(u)du\right),$$

with Q_n *as defined in equation* (2.16).

Proof. We demonstrate briefly that the proof of Masry (1996*b*, proof of Theorem 3) can be carried over to our regression model and assumptions.

The proof employs a small-block and large-block argument. Partition the set $\{1,\dots,n\}$ into $2k+1$ subsets with large blocks of size $l = l_n$ and small blocks of size $s = s_n$, where $s_n/l_n \to 0$, and put $k = k_n = \left\lfloor \frac{n}{l+s} \right\rfloor$.

Now denote $\widetilde{Z}_{n,i} = \sqrt{h_n}Z_{n,i}$, with $Z_{n,i}$ as in equation (2.15), and define large blocks of $\widetilde{Z}_{n,i}$ each having a size of l

$$\eta_j := \sum_{i=j(l+s)+1}^{j(l+s)+l} \widetilde{Z}_{n,i},\ \ 0 \le j \le k-1,$$

and small blocks of $\widetilde{Z}_{n,i}$ each having a size of s

$$\xi_j := \sum_{i=j(l+s)+l+1}^{j(l+s)+l+s} \widetilde{Z}_{n,i},\ \ 0 \le j \le k-1,$$

and for the remaining terms define

$$\zeta_k := \sum_{i=k(l+s)+1}^{n} \widetilde{Z}_{n,i}.$$

Now decompose

$$\sqrt{nh_n}Q_n = \frac{1}{\sqrt{n}}\left(\sum_{j=0}^{k-1}\eta_j + \sum_{j=0}^{k-1}\xi_j + \zeta_k\right) =: \frac{1}{\sqrt{n}}\left(Q_n^{'} + Q_n^{''} + Q_n^{'''}\right).$$

By verification of moments using Lemma 2.18, one can show that the small blocks $Q_n^{''}$ as well as the remaining term $Q_n^{'''}$ are asymptotically negligible if s_n and l_n are chosen in the right way: We need that $l_n/n \underset{n\to\infty}{\longrightarrow} 0, l_n/\sqrt{nh_n} \underset{n\to\infty}{\longrightarrow} 0$, and $n/l_n\alpha(s_n) \underset{n\to\infty}{\longrightarrow} 0$. The existence of such block sizes is guaranteed by the condition on the strong mixing coefficients $\alpha_X(\cdot)$. For further details see Masry (1996*b*, proof of Theorem 3).

For the remaining sum term $Q_n^{'}$, we show that

(i) $\frac{1}{n}\sum\limits_{j=0}^{k-1}\mathbb{E}\left(\eta_j^2\right) \underset{n\to\infty}{\longrightarrow} \theta^2(x)$, with $\theta^2(x) := \sigma_U^2\sigma^2(x)f_X(x)\int A^2(z)dz$,

(ii) $\frac{1}{n}\sum\limits_{j=0}^{k-1}\mathbb{E}\left(\eta_j^2 1_{\{|\eta_j|\geq\varepsilon\theta(x)\sqrt{n}\}}\right) \underset{n\to\infty}{\longrightarrow} 0$, for every $\varepsilon > 0$,

which completes the proof in the manner of Masry (1996*b*, proof of Theorem 3).

Condition (i) is a direct consequence of the proof of Lemma 2.18, which provides that for every summand η_j we have by stationarity

$$\begin{aligned}\mathrm{Var}\left(\eta_j\right) = \mathrm{Var}\left(\eta_0\right) =& l\,\mathrm{Var}(\widetilde{Z}_{n,1}) + 2l\sum_{m=2}^{l}\left(1 - \frac{m-1}{l}\right)\mathrm{Cov}(\widetilde{Z}_{n,1}, \widetilde{Z}_{n,m})\\ =& l\,\theta^2(x)(1 + O(h_n)),\end{aligned}$$

which implies

$$\frac{1}{n}\sum_{j=0}^{k-1}\mathbb{E}\left(\eta_j^2\right) = \frac{1}{n}\sum_{j=0}^{k-1}\mathrm{Var}\left(\eta_j\right) = \frac{k_n l_n}{n}\theta^2(x)(1 + o(1)) \underset{n\to\infty}{\longrightarrow} \theta^2(x),$$

by the choice of the block sizes l_n, s_n, and k_n.

Condition (ii) needs a truncation argument, which differs a bit from the one used by Masry. Here, we put $\widetilde{Z}_{n,i}^t := \widetilde{Z}_{n,i}1_{\{|U_i|\leq t\}}$, where t is a fixed truncation point. It holds $|\widetilde{Z}_{n,i}^t| = O(1/\sqrt{h_n})$. As demonstrated by Masry, this implies that for the truncated sum $\eta_j^t := \sum\limits_{i=j(r+s)}^{j(r+s)+r-1}\widetilde{Z}_{n,i}^t$ the set $1_{\{|\eta_j^t|\geq\varepsilon\theta(x)\sqrt{n}\}}$ becomes an empty set when n is sufficiently large. Thus Condition (ii) holds true for the truncated sum $Q_n^{'t} := \frac{1}{\sqrt{n}}\sum_{j=0}^{k-1}\eta_j^t$.

It remains to show that the truncation is asymptotically negligible. We show that $nh_n\mathrm{Var}(Q_n - Q_n^t) \to 0$, as first $n \to \infty$ and then $t \to \infty$.

We have

$$
\begin{aligned}
&nh_n \mathrm{Var}(Q_n - Q_n^t) \\
&= nh_n \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Z_{n,i} 1_{\{|U_i|>t\}}\right) \\
&= h_n \mathrm{Var}\left(Z_{n,1} 1_{\{|U_1|>t\}}\right) + 2h_n \sum_{k=2}^{n}\left(1-\frac{k-1}{n}\right) \mathrm{Cov}(Z_{n,1}1_{\{|U_1|>t\}}, Z_{n,k}1_{\{|U_k|>t\}}).
\end{aligned}
$$

Note that the process $(U_t 1_{\{|U_t|>t\}})_{t\in\mathbb{Z}}$ is alpha-mixing with the same α-mixing coefficients as $(U_t)_{t\in\mathbb{Z}}$. Similar calculations as in the proof of Lemma 2.18 thus provide

$$
2\, h_n \sum_{k=2}^{n}\left(1-\frac{k-1}{n}\right) \mathrm{Cov}\left(Z_{n,1}1_{\{|U_1|>t\}}, Z_{n,k}1_{\{|U_k|>t\}}\right) = O(h_n),
$$

and

$$
\begin{aligned}
&h_n \mathrm{Var}\left(Z_{n,1}1_{\{|U_1|>t\}}\right) \\
&= \mathbb{E}\left(U_1^2 1_{\{|U_1|>t\}}\right) \mathbb{E}(h_n \sigma^2(X_i) A_{h_n}^2(X_i - x)) \\
&\quad - h_n \left(\mathbb{E}\left(U_1 1_{\{|U_1|>t\}}\right)\right)^2 \left(\mathbb{E}(\sigma(X_i) A_{h_n}(X_i - x))\right)^2 \\
&= \mathbb{E}\left(U_1^2 1_{\{|U_1|>t\}}\right)\left(\sigma^2(x) f_X(x) \int A^2(z)dz\right)(1+o(1)) + \left(\mathbb{E}(U_1 1_{\{|U_1|>t\}})\right)^2 O(h_n),
\end{aligned}
$$

since $\mathbb{E}(h_n\sigma^2(X_i)A_{h_n}^2(X_i - x)) = \left(\sigma^2(x)f_X(x)\int A^2(z)dz\right)(1 + o(1))$ by the proof of Lemma 2.18, and $\mathbb{E}(\sigma(X_i)A_{h_n}(X_i - x)) = O\,(1)$ by that of Lemma 2.9 .

Hence

$$
\lim_{n\to\infty} nh_n \mathrm{Var}(Q_n - Q_n^t) = \mathbb{E}\left(U_1^2 1_{\{|U_1|>t\}}\right)\left(\sigma^2(x) f_X(x) \int A^2(z)dz\right).
$$

Since the third moments of $(U_t)_{t\in\mathbb{Z}}$ are finite by Assumption 1 (iv), Markov's and Hoelder's inequalities provide

$$
\mathbb{E}(U_1^2 1_{\{|U_1|>t\}}) \le (\mathbb{E}|U_1|^3)^{2/3} (\mathbb{E}(1_{\{|U_1|>t\}}))^{1/3} \le \mathbb{E}(|U_1|^3)^{2/3}\left(\frac{\mathbb{E}|U_1|^3}{t^3}\right)^{1/3} \le \frac{\mathbb{E}|U_1|^3}{t},
$$

and thus $nh_n\mathrm{Var}(Q_n - Q_n^t) \to 0$ for $n \to \infty$ and $t \to \infty$. This yields condition (ii) in the manner of Masry (1996*b*, proof of Theorem 3). □

Remark 2.21. *Note that if $h_n = O\,(n^{-\gamma})$, and $\alpha_X(n) = O\left(n^{-\gamma'}\right)$, we have*

$$
\sqrt{n/h_n}\, \alpha_X\left(\sqrt{nh_n}\right) = O\left(n^{\left(\frac{1+\gamma}{2} + \gamma' \frac{1-\gamma}{2}\right)}\right).
$$

Hence for $\gamma' > \frac{1+\gamma}{1-\gamma}$*, we have* $\sqrt{n/h_n}\,\alpha_X\left(\sqrt{nh_n}\right) \to 0$ *as required.*

Lemma 2.22 (Joint asymptotic normality of the local polynomial estimator and its derivatives)**.** *With the same assumptions as in Lemma 2.20, we have*

$$\sqrt{nh_n}\left(\widehat{\boldsymbol{\beta}}(x) - \boldsymbol{\beta}(x) - h_n^{p+1}\mathbf{M}^{-1}\widetilde{\mathbf{B}}\,\frac{m^{(p+1)}(x)}{(p+1)!}\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma_U^2\,\sigma^2(x)/f_X(x)\mathbf{M}^{-1}\boldsymbol{\Gamma}\mathbf{M}^{-1}\right).$$

Proof. We first show that

$$\sqrt{nh_n}\mathbf{X}^\top\mathbf{W}\boldsymbol{\sigma}\mathbf{U} \xrightarrow{D} \mathcal{N}(0, \sigma_U^2\,\sigma^2(x)f_X(x)\boldsymbol{\Gamma}) \tag{2.18}$$

From Lemma 2.20 we have for $a_m := \delta_k(m), m = 0, \dots, p$ that

$$\sqrt{nh_n}t_k \underset{n\to\infty}{\longrightarrow} l_k \sim \mathcal{N}\left(0, \sigma_U^2\sigma^2(x)f_X(x)\int u^{2k}K^2(u)\,du\right),$$

and further for an arbitrary vector $(a_0, \dots, a_p) \in \mathbb{R}^{p+1}$,

$$\sqrt{nh_n}\sum_{k=0}^{p} a_k t_k \xrightarrow{D} \sum_{k=0}^{p} a_k l_k \sim \mathcal{N}\left(0, \sigma_U^2\sigma^2(x)f_X(x)\int A^2(u)du\right).$$

The Cramér-Wold lemma (Cramér & Wold 1936) thus provides

$$\sqrt{nh_n}\mathbf{X}^\top\mathbf{W}\boldsymbol{\sigma}\mathbf{U} \xrightarrow{D} \mathbf{L},$$

with $\mathbf{L} := (l_0, \dots, l_p)^\top \sim \mathcal{N}(0, \sigma_U^2\,\sigma^2(x)f_X(x)\boldsymbol{\Gamma})$, which yields equation (2.18).

Corollary 2.10 gives $(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1} \xrightarrow{\mathrm{P}} 1/f_X(x)\mathbf{M}^{-1}$. From Slutsky's theorem and the fact that $Y \sim \mathcal{N}(a, \boldsymbol{\Sigma}) \Rightarrow \mathbf{C}Y \sim \mathcal{N}(\mathbf{C}a, \mathbf{C}^\top\boldsymbol{\Sigma}\mathbf{C})$ for an arbitrary matrix $\mathbf{C}$, we therefore obtain that

$$\sqrt{nh_n}(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\boldsymbol{\sigma}\mathbf{U} \xrightarrow{D} 1/f_X(x)\mathbf{M}^{-1}\mathbf{L} \sim \mathcal{N}\left(0, \sigma_U^2\,\sigma^2(x)/f_X(x)\mathbf{M}^{-1}\boldsymbol{\Gamma}\mathbf{M}^{-1}\right).$$

Since

$$\widehat{\boldsymbol{\beta}}(x) = \boldsymbol{\beta}(x) + \mathbf{B}_x + (\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\boldsymbol{\sigma}\mathbf{U} + o_P\left(h_n^{p+1}\right)$$

and $\mathbf{B}_x \xrightarrow{\mathrm{P}} h_n^{p+1}\frac{m^{(p+1)}(x)}{(p+1)!}\mathbf{M}^{-1}\widetilde{\mathbf{B}}$, Slutsky's theorem yields the assertion. □

The convergence of the local polynomial estimator $\widehat{m}(x)$ is now a direct consequence of the previous results:

Corollary 2.23 (Asymptotic normality of the local polynomial estimator)**.** *With the same assump-*

tions as in Lemma 2.20, we have

$$\sqrt{nh_n}\left(\widehat{m}(x)-m(x)-h_n^{p+1}\frac{m^{(p+1)}(x)}{(p+1)!}\mathbf{e}_1^{\mathsf{T}}\mathbf{M}^{-1}\widetilde{\mathbf{B}}\right)$$
$$\xrightarrow{D}\mathcal{N}\left(0,\sigma_U^2\,\sigma^2(x)/f_X(x)\left[\mathbf{M}^{-1}\mathbf{\Gamma}\mathbf{M}^{-1}\right]_{1,1}\right).$$

Remark 2.24 (Assumption on the mixing coefficient)**.** *If we assume exponential decreasing strong mixing coefficients for* $(X_t)_{t\in\mathbb{Z}}$*, i.e.* $\alpha_X(k)\leq\exp(-C_\alpha k)$*, then the joint normality as well as the uniform strong consistency of the matrix* $\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X}$ *will hold true for any choice of the bandwidth* h_n *with* $h_n\underset{n\to\infty}{\longrightarrow}0$, $nh_n/\log(n)\underset{n\to\infty}{\longrightarrow}\infty$.

Chapter 3

A more efficient local polynomial estimator

In this chapter we derive a modification of the local polynomial estimator based on a prewhitening transformation of the data. Calculating its limiting distribution, we show that this estimator is more efficient than the conventional one if the observation error originates from a strongly mixing process.

Contents

3.1 Estimation method

3.1.1 Estimation idea and an infeasible efficient estimator

We consider the homoscedastic regression model

$$Y_t = m(X_t) + U_t, \quad t = 1, \dots, n$$

as the special case $\sigma^2(\cdot) \equiv 1$ of the model introduced in the previous chapter. The general heteroscedastic case will be discussed in Chapter 4.

We derive a more efficient estimator by "prewhitening" the original regression model in a way that the filtered data set is less correlated. For this purpose, we consider a linear approximation of the error term U_t of order q

$$U_t \approx \alpha_1 U_{t-1} + \ldots + \alpha_q U_{t-q} =: P_q U_t,$$

with $\alpha_1, \ldots, \alpha_q$ chosen in a way to minimize the mean squared error $\mathbb{E}(U_t - P_q U_t)^2$. The resulting linear combination $P_q U_t$ is the orthogonal projection of U_t on the linear space spanned by $U_{t-1}, \ldots, U_{t-q}$.

Define the filtered series

$$\overline{Y_t} = Y_t - P_q U_t,$$

and further $\widehat{m}(x)$ and $\overline{m}(x)$ the pth order local polynomial estimators of $m(x)$ on $(Y_t)_{t=1}^n$ and $(\overline{Y_t})_{t=1}^n$ respectively. Note that if the error process was an AR(q)- process:

$$U_t = \alpha_1 U_{t-1} + \ldots + \alpha_q U_{t-q} + \varepsilon_t, \text{ with } \varepsilon_t \text{ i.i.d.},$$

the filtering procedure would give $\overline{Y_t} = m(X_t) + \varepsilon_t$. Since $\mathrm{Var}(\varepsilon_t) \leq \mathrm{Var}(U_t)$, we expect a more efficient regression on the filtered data. We will now see that even if the linear approximation is far from matching the true autocorrelation structure, the filtering procedure still leads to a more efficient estimation for almost all error processes that possess serial correlation.

Define the prewhitened errors $\overline{U}_t := U_t - P_q U_t$, and the error variances $\sigma_U^2 := \mathbb{E}(U_t^2)$, $\sigma_{\overline{U}}^2 := \mathbb{E}(\overline{U_t}^2)$. If the prewhitened error process $(\overline{U_t})_{t\in\mathbb{Z}}$ is $\alpha-$mixing with mixing coefficients that satisfy Assumption 1 (iv), then Corollary 2.23 consequently provides the following convergence results for the estimators on $(Y_t)_{t=1}^n$ and $(\overline{Y_t})_{t=1}^n$ respectively:

$$\boxed{\begin{aligned}\sqrt{nh_n}\left(\widehat{m}(x) - m(x) - h_n^{p+1}B(x)\right) &\xrightarrow{D} \mathcal{N}\left(0, \sigma_U^2\, V(x)\right),\\ \sqrt{nh_n}\left(\overline{m}(x) - m(x) - h_n^{p+1}B(x)\right) &\xrightarrow{D} \mathcal{N}\left(0, \sigma_{\overline{U}}^2\, V(x)\right).\end{aligned}}$$

with $B(x) := \left[\frac{m^{(p+1)}(x)}{(p+1)!}\mathbf{M}^{-1}\widetilde{\mathbf{B}}\right]_{1,1}$ the bias term, and $V(x) := \frac{1}{f_X(x)}\left[\mathbf{M}^{-1}\mathbf{\Gamma}\mathbf{M}^{-1}\right]_{1,1}$ the variance term.

Furthermore, by the projection property of $P_q U_t$, it holds for the variances of the original and the filtered error process respectively,

$$\begin{aligned}
\sigma_U^2 &= \mathbb{E}(U_t - P_q U_t + P_q U_t)^2 \\
&= \mathbb{E}\left(\overline{U}_t\right)^2 + \mathbb{E}(P_q U_t)^2 + 2\,\mathbb{E}((U_t - P_q U_t) P_q U_t) \\
&= \mathbb{E}\left(\overline{U}_t\right)^2 + \mathbb{E}(P_q U_t)^2 + 0 \\
&\geq \mathbb{E}\left(\overline{U}_t\right)^2 = \sigma_{\overline{U}}^2.
\end{aligned}$$

If the filtered error process actually differs from the original error process i.e., $\mathrm{P}(0 \neq P_q U_t) > 0$, we have[1]

$$\sigma_U^2 > \sigma_{\overline{U}}^2.$$

The previous calculations yield that the estimator $\overline{m}(x)$ has an asymptotically smaller variance than $\widehat{m}(x)$ and the same bias. It should hence be preferred.

Remark 3.1 (Theoretical asymptotic efficiency gain)**.** *The theoretical asymptotic efficiency of the proposed estimator relative to the conventional one in purely variance terms is $\sigma_{\overline{U}}^2/\sigma_U^2$, which doesn't exceed one by the previous calculations.*

For a nonzero bias, comparing the relative efficiency at the respectively optimal bandwidths would give us the slightly smaller value

$$\frac{MSE(\overline{m}(x))}{MSE(\widehat{m}(x))} = \left(\frac{\sigma_{\overline{U}}^2}{\sigma_U^2}\right)^{\frac{2p+2}{2p+3}}.$$

This is due to the fact, that the optimal bandwidth for the proposed estimator is the same as for the conventional one, but with $(\sigma_{\overline{U}}^2)^{1/(2p+3)}$ instead of $(\sigma_U^2)^{1/(2p+3)}$. The resulting MSE at the respectively optimal bandwidths is therefore distinct by the factors $(\sigma_U^2)^{(2p+2)/(2p+3)}$ and $(\sigma_{\overline{U}}^2)^{(2p+2)/(2p+3)}$, respectively. In Chapter 5 we will comment the optimal bandwidth choice in more detail.

Example 3.2 (Efficiency gain for AR(2)-error process)**.** Suppose that the error process is of AR(2)-type

$$U_t = \alpha_1 U_{t-1} + \alpha_2 U_{t-2} + \varepsilon_t, \qquad t \in \mathbb{Z}, \tag{3.1}$$

with $(\varepsilon_t)_{t\in\mathbb{Z}}$ an i.i.d. process fulfilling $\mathbb{E}(\varepsilon_t) = 0$ and $\mathrm{Var}(\varepsilon_t) = \sigma_\varepsilon^2$, and α_1, α_2 such that $1 - \alpha_1 z - \alpha_2 z^2 \neq 0$, for all $z \in \mathbb{C}$, $|z| < 1$.[2]

[1]For an AR(q) error process with $\mathrm{Var}(\varepsilon_t) > 0$, the assertion $\mathrm{P}(P_q U_t = 0) = 1$ is equivalent to the case of independent errors: $\alpha_i = 0$ for all $i = 1, \ldots, q$.

[2]This assumption assures that the process is stationary, see for example Brockwell & Davis (1990).

The variance σ_U^2 of such a process can be calculated as follows: Squaring equation (3.1) and taking expectations, we find that the variance must satisfy:

$$\sigma_U^2 = \alpha_1^2\sigma_U^2 + \alpha_2^2\sigma_U^2 + 2\alpha_1\alpha_2\mathrm{Cov}(U_0, U_1) + \sigma_\varepsilon^2, \qquad t \in \mathbb{Z}, \tag{3.2}$$

where we used that the process is stationary and satisfies $\mathrm{Cov}(\varepsilon_t, U_{t-1}) = 0$.

Further, multiplying equation (3.1) by U_{t-1} and taking expectations gives

$$\mathrm{Cov}(U_0, U_1) = \alpha_1\sigma_U^2 + \alpha_2\mathrm{Cov}(U_0, U_1),$$

or equivalently $\mathrm{Cov}(U_0, U_1) = \alpha_1/(1-\alpha_2)\sigma_U^2$. Using this expression in equation (3.2) results in the following formula for the variance:

$$\sigma_U^2 = \frac{1-\alpha_2}{(1+\alpha_2)((1-\alpha_2)^2 - \alpha_1^2)}\,\sigma_\varepsilon^2,$$

In contrast to that, the filtering procedure with $q = 2$ would lead to $\overline{U}_t = \varepsilon_t$, $t \in \mathbb{Z}$ with $\sigma_{\overline{U}}^2 = \sigma_\varepsilon^2$. If we would prewhiten this error process with an AR(2) process, the relative efficiency of the proposed estimator relative to the conventional one in pure terms of variance would hence be

$$\frac{\mathrm{Var}(\widetilde{m}(x))}{\mathrm{Var}(\widehat{m}(x))} = \frac{\sigma_{\overline{U}}^2}{\sigma_U^2} = \frac{(1+\alpha_2)((1-\alpha_2)^2 - \alpha_1^2)}{1-\alpha_2} = 1 - \alpha_2^2 - \alpha_1^2\left(\frac{1+\alpha_2}{1-\alpha_2}\right) \leq 1 - 0,$$

which is strictly smaller than one except for the case of independent errors: $\alpha_1 = \alpha_2 = 0$. In the case that either $\alpha_1 = 0$ or $\alpha_2 = 0$, we would obtain a relative efficiency of $1-\alpha_2^2$ or $1-\alpha_1^2$ respectively. Note that both expressions tend to zero as $\alpha_2 \to 1$ and $\alpha_1 \to 1$ respectively.

As we commented in the previous remark, the efficiency gain measured in MSE in this example would be $\left(1 - \alpha_2^2 - \alpha_1^2\left(\frac{1+\alpha_2}{1-\alpha_2}\right)\right)^{\frac{2p+2}{2p+3}}$. ◇

Note that $\overline{m}(x)$ is an *oracle*-type estimator since its definition contains knowledge that only an oracle could know. In the next section we replace the unknown quantities by their estimates in order to obtain a feasible estimation.

3.1.2 The proposed feasible estimator

In practice, the error terms U_t and the parameter vector $\boldsymbol{\alpha} := (\alpha_1, \dots, \alpha_q)^\mathsf{T}$ are unknown. Hence, the estimator $\overline{m}(x)$ is infeasible. We thus replace the unknown quantities U_t and $\boldsymbol{\alpha}$ by their estimates $\widehat{U}_t$ and $\widehat{\boldsymbol{\alpha}}$ respectively to obtain a feasible estimator $\widetilde{m}(x)$. For this purpose, we conduct a previous pth order local polynomial smoothing that gives an estimate of the error terms. The parameter vector is then obtained by least squares estimation on the residuals of this prior smoothing. The resulting estimation procedure is as follows:

Estimation Procedure (Proposed estimator in the homoscedastic case)**.**

1. *Pilot fit:* Obtain a preliminary local pth order polynomial smoothing Y_t on X_t for $t = 1, \ldots, n$, using kernel $K_0(\cdot)$ and bandwidth $h_0 = h_0(n)$. Denote the estimates $\widehat{m}(X_t)$ and calculate an approximation of the residuals U_t by

$$\widehat{U}_t := Y_t - \widehat{m}(X_t), \quad t = 1, \ldots, n.$$

2. *Calculation of filter parameters:* Define $V_t := U_t - \sum_{k=1}^{q} \alpha_k U_{t-k}$. Having the linear (auto-) regression model

$$\begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix} = \begin{pmatrix} U_0 & \ldots & U_{1-q} \\ \vdots & \ddots & \vdots \\ U_{n-1} & \ldots & U_{n-q} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_q \end{pmatrix} + \begin{pmatrix} V_1 \\ \vdots \\ V_n \end{pmatrix} =: \mathbf{U}_q \, \boldsymbol{\alpha} + \mathbf{V}$$

in mind, we calculate a least squares estimate $\widehat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ by

$$\widehat{\boldsymbol{\alpha}} := \left(\widehat{\mathbf{U}}_q^{\mathsf{T}} \widehat{\mathbf{U}}_q \right)^{-1} \widehat{\mathbf{U}}_q^{\mathsf{T}} \widehat{\mathbf{U}},$$

with $\widehat{\mathbf{U}} = (\widehat{U}_1, \ldots, \widehat{U}_n)^{\mathsf{T}}$, and $\widehat{\mathbf{U}}_q$ defined like $\mathbf{U}_q$ but with U_t replaced by $\widehat{U}_t$.

3. *Pre-Whitening:* Calculate an approximation $\widetilde{\mathbf{Y}} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_n)$ of the filtered series $\overline{\mathbf{Y}}$ by

$$\widetilde{Y}_t := Y_t - \widehat{\alpha}_1 \widehat{U}_{t-1} - \ldots - \widehat{\alpha}_q \widehat{U}_{t-q}, \quad t = 1, \ldots, n.$$

4. *Final fit:* The proposed estimator $\widetilde{m}(x)$ is obtained by pth order local polynomial smoothing of $\widetilde{Y}_t$ on X_t, using kernel $K(\cdot)$ and bandwidth $h = h(n)$.

3.2 Efficiency of the proposed estimator

We prove that our proposed estimator is more efficient than the conventional one by showing the following convergence result:

$$\sqrt{nh_n} \left(\widetilde{m}(x) - m(x) - h_n^{p+1} B(x) \right) \xrightarrow{D} \mathcal{N} \left(0, \sigma_U^2 \, V(x) \right). \tag{3.3}$$

For this purpose, we show that $\overline{m}(x)$ and $\widetilde{m}(x)$ are asymptotically equivalent. To facilitate the analysis, we suppose Assumption 1 to hold and add the following assumptions:

Assumption 2.

(i) The stationary process $(X_t)_{t\in\mathbb{Z}} \in D$, with D a compact interval, is strongly mixing with exponentially decreasing mixing coefficients, i.e. $\alpha(k) \le \exp(-C_\alpha k)$.

(ii) The stationary error process $(U_t)_{t\in\mathbb{Z}}$ has finite sixth moments, is independent of $(X_t)_{t\in\mathbb{Z}}$ and strongly mixing with mixing coefficients $\alpha_U(\cdot)$ that satisfy $\sum\limits_{k=0}^{\infty}(k^2+1)\,\alpha_U(k)^{1/3} < \infty$.

(iii) The kernels $K_0(\cdot)$ and $K(\cdot)$ are Lipschitz-continuous, non-negative, strictly positive in zero, bounded, and have support on $[-1,1]$.

(iv) The bandwidths h_0 and h satisfy $h_0 \underset{n\to\infty}{\longrightarrow} 0$, $h \underset{n\to\infty}{\longrightarrow} 0$, $nh_0/\log(n) \underset{n\to\infty}{\longrightarrow} \infty$, $nh \underset{n\to\infty}{\longrightarrow} \infty$, $\sqrt{h}/(\sqrt{n}h_0) \underset{n\to\infty}{\longrightarrow} 0$ and $\sqrt{nh}h_0^{p+1} \underset{n\to\infty}{\longrightarrow} 0$.

Remark 3.3 (Some notes on the assumptions)**.** *Under the mixing conditions of Assumption 2 (i), the temporal dependence among $(X_t)_{t\in\mathbb{Z}}$ decreases sufficiently fast so that the uniform strong convergence result for the matrix $\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}$ as stated in Lemma 2.13, as well as the joint asymptotic normality provided in Corollary 2.23 hold true for any choice of the bandwidths. If we assume a certain order of convergence on the bandwidths, i.e. $h_0 = O\left(n^{-\gamma_0}\right)$ and $h = O\left(n^{-\gamma_1}\right)$, the assumptions on the strong mixing coefficients can be weakened to $\alpha_X(n) = O\left(n^{-\eta}\right)$ for a suitable choice of $\eta > 0$.*[3]

We will need the error process to have finite sixth moments and summable mixing coefficients as stated in Assumption 2 (ii) in order to derive asymptotic results for variance type terms. Assumption 2 (iii) is similar to previous assumptions.

One important reason for our procedure to work, is that the bias from the first estimation can be made asymptotically negligible by undersmoothing in the first stage. This is guaranteed by Assumption 2 (iv) on the bandwidths ($\sqrt{nh}h_0^{p+1} \underset{n\to\infty}{\longrightarrow} 0$). The requirement $nh_0/\log(n) \underset{n\to\infty}{\longrightarrow} \infty$ is needed for the uniform convergence of $\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}$ in the first estimation, see Lemma 2.13.

We show that h_0 and h can be chosen in a way that Assumption 2 (iv) holds true. A useful choice of the main bandwidth h is to minimize the mean squared error (MSE) of $\widetilde{m}(x)$. Corollary 3.5 provides that the leading bias and variance terms have the asymptotic orders $B(x) = O_P\left(h^{p+1}\right)$ and $V(x) = O_P\left(\frac{1}{nh}\right)$ respectively. Balancing those terms in order to minimize the asymptotic MSE we obtain for the order of the optimal bandwidth

$$h \asymp n^{-\frac{1}{2p+3}}.$$

[3]Remark 2.21 provides $\eta > \max\left\{\frac{1+\gamma_0}{1-\gamma_0}, \frac{1+\gamma_1}{1-\gamma_1}\right\}$ for asymptotic normality, and $\eta > \max\left\{\frac{10}{1-\gamma_0}, \frac{10}{1-\gamma_1}\right\}$ for uniform strong convergence of $\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}$.

With this choice of h, the assumptions on the bandwidths are satisfied if we choose the prior bandwidth h_0 of slightly smaller order

$$h_0 \asymp h^{\delta}, \text{ with } 1 < \delta < 3.$$

We show that h_0 and h satisfy all bandwidth conditions:

- $h_0 \underset{n\to\infty}{\longrightarrow} 0$, $h \underset{n\to\infty}{\longrightarrow} 0$.
- $nh_0/\log(n) \asymp n^{\frac{2p+3-\delta-\gamma}{2p+3}} \underset{n\to\infty}{\longrightarrow} \infty$, *since* $p \geq 1$, *and* $\delta < 3$.[4]
- $nh \asymp n^{\frac{2p+2}{2p+3}} \underset{n\to\infty}{\longrightarrow} \infty$.
- $\sqrt{h}/(\sqrt{n}h_0) \asymp n^{-1/(4p+6)}n^{-1/2}n^{\delta/(2p+3)} = n^{\frac{-1-(2p+3)+2\delta}{4p+6}} \underset{n\to\infty}{\longrightarrow} 0$, *since* $p \geq 1$ *and* $\delta < 3$.
- $\sqrt{nh}{h_0}^{p+1} \asymp n^{1/2}n^{-1/(4p+6)}n^{-(p+1)\delta/(2p+3)} = n^{\frac{(2p+2)(1-\delta)}{4p+6}} \underset{n\to\infty}{\longrightarrow} 0$, *since* $\delta > 1$.

We can now formulate our main result which directly implies equation (3.3).

Theorem 3.4. *Under Assumptions 1 and 2 it holds*

$$\widetilde{m}(x) = \overline{m}(x) + o_P\left(\frac{1}{\sqrt{nh}}\right) + O_P\left(h_0^{p+1}\right).$$

Corollary 3.5. *It holds*

$$\sqrt{nh}\left(\widetilde{m}(x) - m(x) - h^{p+1}B(x)\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma_U^2\, V(x)\right).$$

Proof. We have

$$\begin{aligned}
&\sqrt{nh}\left(\widetilde{m}(x) - m(x) - h^{p+1}\frac{m^{(p+1)}(x)}{(p+1)!}\mathbf{M}^{-1}\widetilde{\mathbf{B}}\right) \\
&= \sqrt{nh}\left(\overline{m}(x) + o_P\left(\frac{1}{\sqrt{nh}}\right) + O_P(h_0^{p+1}) - m(x) - h^{p+1}\frac{m^{(p+1)}(x)}{(p+1)!}\mathbf{M}^{-1}\widetilde{\mathbf{B}}\right) \\
&= \sqrt{nh}\left(\overline{m}(x) - m(x) - h^{p+1}\frac{m^{(p+1)}(x)}{(p+1)!}\mathbf{M}^{-1}\widetilde{\mathbf{B}}\right) + o_P(1) + O_P\left(\sqrt{nh}h_0^{p+1}\right).
\end{aligned}$$

Undersmoothing (Assumption 2 (iv)) provides $\sqrt{nh}h_0^{p+1} = o_P(1)$. Slutsky's theorem yields the assertion. $\square$

[4]Note that $\log(n) = o\left(n^{\gamma}\right)$ for every $\gamma > 0$.

A combination of the previous results gives

$$\boxed{\begin{aligned}\sqrt{nh}\left(\widehat{m}(x) - m(x) - h^{p+1}B(x)\right) &\xrightarrow{D} \mathcal{N}\left(0, \sigma_U^2\, V(x)\right),\\ \sqrt{nh}\left(\widetilde{m}(x) - m(x) - h^{p+1}B(x)\right) &\xrightarrow{D} \mathcal{N}\left(0, \sigma_{\overline{U}}^2\, V(x)\right).\end{aligned}}$$

The proposed estimator has thus an asymptotically smaller variance than the conventional one but the same bias and should hence be preferred.

It remains to prove Theorem 3.4:

Proof of Theorem 3.4. Keep in mind that

$$\widetilde{m}(x) = \mathbf{e}_1^\mathsf{T}\left((\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\widetilde{\mathbf{Y}}\right).$$

It holds

$$\begin{aligned}\widetilde{Y}_t &= Y_t - \sum_{k=1}^q \widehat{\alpha}_k \widehat{U}_{t-k}\\ &= Y_t - \sum_{k=1}^q \alpha_k U_{t-k} - \sum_{k=1}^q (\widehat{\alpha}_k - \alpha_k)U_{t-k} + \sum_{k=1}^q \alpha_k (U_{t-k} - \widehat{U}_{t-k})\\ &\quad + \sum_{k=1}^q (\widehat{\alpha}_k - \alpha_k)(U_{t-k} - \widehat{U}_{t-k})\\ &= \overline{Y_t} - \sum_{k=1}^q (\widehat{\alpha}_k - \alpha_k)U_{t-k} + \sum_{k=1}^q \alpha_k(\widehat{m}(X_{t-k}) - m(X_{t-k}))\\ &\quad + \sum_{k=1}^q (\widehat{\alpha}_k - \alpha_k)(\widehat{m}(X_{t-k}) - m(X_{t-k})).\end{aligned}$$

Therefore we have

$$\mathbf{X}^\mathsf{T}\mathbf{W}\widetilde{\mathbf{Y}} = \mathbf{X}^\mathsf{T}\mathbf{W}\overline{\mathbf{Y}} - \mathbf{R}_{n,1} + \mathbf{R}_{n,2} + \mathbf{R}_{n,3},$$

with

$$\mathbf{R}_{n,1} := \frac{1}{nh}\begin{pmatrix} \sum_{t=1}^n K\left(\frac{X_t - x}{h}\right) \sum_{k=1}^q (\widehat{\alpha}_k - \alpha_k)U_{t-k} \\ \vdots \\ \sum_{t=1}^n K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^p \sum_{k=1}^q (\widehat{\alpha}_k - \alpha_k)U_{t-k}\end{pmatrix},$$

$$\mathbf{R}_{n,2} := \frac{1}{nh}\begin{pmatrix} \sum_{t=1}^{n} K\left(\frac{X_t-x}{h}\right)\sum_{k=1}^{q}\alpha_k(\widehat{m}(X_{t-k})-m(X_{t-k})) \\ \vdots \\ \sum_{t=1}^{n} K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^p\sum_{k=1}^{q}\alpha_k(\widehat{m}(X_{t-k})-m(X_{t-k})) \end{pmatrix},$$

$$\mathbf{R}_{n,3} := \frac{1}{nh}\begin{pmatrix} \sum_{t=1}^{n} K\left(\frac{X_t-x}{h}\right)\sum_{k=1}^{q}(\alpha_k-\widehat{\alpha}_k)(\widehat{m}(X_{t-k})-m(X_{t-k})) \\ \vdots \\ \sum_{t=1}^{n} K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^p\sum_{k=1}^{q}(\alpha_k-\widehat{\alpha}_k)(\widehat{m}(X_{t-k})-m(X_{t-k})) \end{pmatrix}.$$

In Lemmas 3.10 - 3.12 we show that $\mathbf{R}_{n,i} = O_P\left(h_0^{p+1}\right) + o_P\left(\frac{1}{\sqrt{nh}}\right), i = 1, 2, 3$. For this purpose, we first show in Lemma 3.9 that the filtering parameter $\boldsymbol{\alpha}$ can be estimated with a rate of $O_P\left(h_0^{p+1} + \frac{1}{nh_0} + \frac{1}{\sqrt{n}}\right)$, that is sufficiently small for our needs thanks to the assumptions on the bandwidths.

The rates of the remaining terms can be explained as follows: Since we make use of the averaging effect over the error process and since the estimation of $\boldsymbol{\alpha}$ is $o_P\left(\frac{1}{\sqrt{nh}}\right)$ by the assumptions on the bandwidths, we get for the first term that $\mathbf{R}_{n,1} = o_P\left(\frac{1}{nh}\right)$. The second term converges with a rate of $\mathbf{R}_{n,2} = O_P\left(h_0^{p+1}\right) + o_P\left(\frac{1}{\sqrt{nh}}\right)$, since the bias of the prior polynomial fitting is of order $O_P\left(h_0^{p+1}\right)$ and - thanks to averaging - the order variance type term is smaller than that of the original variance term. The rate of $\mathbf{R}_{n,3} = o_P\left(\frac{h_0^{p+1}}{\sqrt{nh}}\right) + o_P\left(\frac{1}{nh}\right)$ follows as a direct combination of the convergence results for $\mathbf{R}_{n,2}$ and the filtering parameter $\boldsymbol{\alpha}$.

Substituting these expressions into $\widetilde{m}(x)$, we have

$$\widetilde{m}(x) = \overline{m}(x) - \mathbf{Q}_{n,1} + \mathbf{Q}_{n,2} + \mathbf{Q}_{n,3},$$

where $\mathbf{Q}_{n,i} := \mathbf{e}_1^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{R}_{n,i},\ i = 1, 2, 3.$

By Corollary 2.10 we have $(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} = (\mathbf{M}f_X(x))^{-1} + o_P(1)$. We therefore get

$$\mathbf{Q}_{n,i} = \mathbf{e}_1^\mathsf{T}(\mathbf{M}f_X(x))^{-1}\mathbf{R}_{n,i}(1 + o_P(1)) = O_P(h_0^{p+1}) + o_P\left(\frac{1}{\sqrt{nh}}\right),\ i = 1, 2, 3,$$

which concludes the proof. □

It remains to study the asymptotic orders of $\mathbf{R}_{n,1}, \mathbf{R}_{n,2}$ and $\mathbf{R}_{n,3}$ and the filtering parameter $\boldsymbol{\alpha}$. We first provide some auxiliary Lemmas that will facilitate our calculations.

Lemma 3.6 (Upper bound for the summed covariances)**.** *Under Assumption 2 (ii), it holds*

$$\sum_{i,j=1}^{n} |\mathbb{E}(U_i U_j)| = O(n).$$

Proof. Since the error process has zero mean and finite third moments, the covariance inequality for strongly mixing processes (Lemma 2.5) provides

$$|\mathbb{E}(U_i U_j)| \leq 4\, \alpha_U(|i-j|)^{1/3} \|U_i\|_3 \|U_j\|_3 \leq C\, \alpha_U(|i-j|)^{1/3}.$$

The summability condition on $\alpha_U(\cdot)^{1/3}$ by Assumption 2 (ii) hence gives

$$\sum_{i,j=1}^{n} |\mathbb{E}(U_i U_j)| \leq C \sum_{i=1}^{n} \sum_{t=0}^{i} \alpha_U(t)^{1/3} \leq n\, C = O(n).$$

$\square$

Lemma 3.7 (Upper bound for the fourth order cumulant)**.** *Under Assumption 2 (ii), it holds with $i \leq j \leq s \leq t$,*

$$|\mathrm{Cum}(U_i, U_j, U_s, U_t)| \leq C(\alpha_U(\max\{j-i, s-j, s-t\}))^{1/3}.$$

Proof. We distinguish the different cases for the maximum distance $T := \max\{j-i, s-j, s-t\}$.

(i) $T = j - i$

From the covariance inequality for strongly mixing processes (Lemma 2.5) we get

$$\begin{aligned}
|\mathbb{E}(U_i U_j U_s U_t)| &= |\mathrm{Cov}(U_i, U_j U_s U_t)| \leq 4\, (\alpha_U(T))^{1/3}\, \|U_i\|_6 \|U_j U_s U_t\|_2, \\
|\mathbb{E}(U_i U_j)| &= \mathrm{Cov}(U_i, U_j) \leq 4\, (\alpha_U(T))^{1/3}\, \|U_i\|_3 \|U_j\|_3, \\
|\mathbb{E}(U_s U_t)| &\leq 4\, (\alpha_U(t-s))^{1/3} \|U_s\|_3 \|U_t\|_3 \leq 4 \|U_s\|_3 \|U_t\|_3.
\end{aligned}$$

With analogous results for $\mathbb{E}(U_i U_s)$, $\mathbb{E}(U_i U_t)$, and $\mathbb{E}(U_j U_s)$, $\mathbb{E}(U_j U_t)$ respectively, we get

$$\begin{aligned}
&|\mathrm{Cum}(U_i, U_j, U_s, U_t)| \\
&\leq |\mathbb{E}(U_i U_j U_s U_t)| + |\mathbb{E}(U_i U_j)||\mathbb{E}(U_s U_t)| \\
&\quad + |\mathbb{E}(U_i U_s)||\mathbb{E}(U_j U_t)| + |\mathbb{E}(U_i U_t)||\mathbb{E}(U_j U_s)| \\
&\leq 4\, (\alpha_U(T))^{1/3} \|U_i\|_6 \|U_j U_s U_t\|_2 + 4\, (\alpha_U(T))^{1/3} \|U_i\|_3 \|U_j\|_3\, 4\, \|U_s\|_3 \|U_t\|_3 \\
&\quad + 4\, (\alpha_U(T))^{1/3} \|U_i\|_3 \|U_s\|_3\, 4\, \|U_j\|_3 \|U_t\|_3 \\
&\quad + 4\, (\alpha_U(T))^{1/3} \|U_i\|_3 \|U_t\|_3\, 4\, \|U_j\|_3 \|U_s\|_3 \\
&\leq C\, (\alpha_U(T))^{1/3},
\end{aligned}$$

since Hoelder's inequality provides $\|U_jU_sU_t\|_2 \leq \|U_j\|_6\|U_s\|_6\|U_t\|_6$, and the sixth moments of the error process are finite.

(ii) $T = s - j$
Again Lemma 2.5 gives

$$|\mathrm{Cov}(U_iU_j, U_sU_t)| \leq 4\,(\alpha_U(T))^{1/3}\|U_iU_j\|_3\|U_sU_t\|_3,$$
$$|\mathbb{E}(U_iU_s)| \leq 4\,(\alpha_U(T))^{1/3}\|U_i\|_3\|U_s\|_3.$$

With similar calculations for the remaining terms we get

$$\begin{aligned}
&|\mathrm{Cum}(U_i, U_j, U_s, U_t)| \\
&\leq |\mathrm{Cov}(U_iU_j, U_sU_t)| + |\mathbb{E}(U_iU_s)||\mathbb{E}(U_jU_t)| + |\mathbb{E}(U_iU_t)||\mathbb{E}(U_jU_s)| \\
&\leq 4\,(\alpha_U(T))^{1/3}\|U_i\|_6\|U_j\|_6\|U_s\|_6\|U_t\|_6 + 32\,(\alpha_U(T))^{2/3}\|U_i\|_3\|U_j\|_3\|U_s\|_3\|U_t\|_3 \\
&\leq C\,(\alpha_U(T))^{1/3}.
\end{aligned}$$

(iii) $T = s - t$
This case is similar to (i).

□

Lemma 3.8 (Regularity of the covariance matrix of a stationary process)**.** *For a stationary process* $(Z_t)_{t\in\mathbb{N}}$ *that satisfies*

- $\mathbb{E}(Z_t) = 0$ *and* $\mathrm{Var}(Z_t) = \sigma_Z^2 \in (0, \infty)$,
- $\mathbb{E}(Z_1Z_t) \underset{t\to\infty}{\longrightarrow} 0$,

the covariance matrix $\mathbf{\Gamma}_n := \mathrm{Cov}(Z_i, Z_j)_{i,j=1,\ldots,n}$ *is regular for all* $n \in \mathbb{N}$.

Proof. Obviously $\mathbf{\Gamma}_1 > 0$. Let $n \in \mathbb{N}$ with $\mathbf{\Gamma}_n$ invertible and $\mathbf{\Gamma}_{n+1}$ not invertible. The singularity of $\mathbf{\Gamma}_{n+1}$ implies the existence of $\widetilde{c} = (\widetilde{c}_1, \ldots, \widetilde{c}_{n+1})^\mathsf{T} \neq (0, \ldots, 0)^\mathsf{T}$ with

$$0 = \widetilde{c}^{\,T}\mathrm{Cov}(Z_1, \ldots, Z_{n+1})\widetilde{c} = \sum_{1\leq k,l\leq n+1} \widetilde{c}_k\widetilde{c}_l\,\mathrm{Cov}(Z_k, Z_l) = \mathbb{E}\left(\sum_{k=1}^{n+1}\widetilde{c}_kZ_k\right)^2,$$

and hence $\sum\limits_{j=1}^{n+1}\widetilde{c}_jZ_j = 0$ almost surely. Regularity of $\mathbf{\Gamma}_n$ implies $\widetilde{c}_{n+1} \neq 0$. Set $c_j := -\widetilde{c}_j/\widetilde{c}_{n+1}$ to get $Z_{n+1} = \sum\limits_{j=1}^{n} c_jZ_j$ almost surely. The stationarity of $(Z_t)_{t\in\mathbb{N}}$ yields

$$Z_{n+2} = \sum_{j=1}^{n-1} c_jZ_{j+1} + c_nZ_{n+1} = \sum_{j=1}^{n-1} c_jZ_{j+1} + \sum_{j=1}^{n} c_jZ_j =: \sum_{j=1}^{n} c_j^{(n+2)}Z_j \text{ almost surely,}$$

and similarly for an arbitrary $m > n$

$$Z_m = \sum_{j=1}^{n} c_j^{(m)} Z_j \text{ almost surely.}$$

Note that the coefficients $c_j^{(m)}$ are bounded, since

$$\begin{aligned}\sigma_Z^2 = \operatorname{Var}(Z_m) = \operatorname{Var}\left(\sum_{j=1}^{n} c_j^{(m)} Z_j\right) &= \left(c_1^{(m)}, \dots, c_n^{(m)}\right) \boldsymbol{\Gamma}_n \begin{pmatrix} c_1^{(m)} \\ \vdots \\ c_n^{(m)} \end{pmatrix} \\ &\geq \lambda_{\min}(\boldsymbol{\Gamma}_n) \sum_{j=1}^{n} \left(c_j^{(m)}\right)^2.\end{aligned}$$

Hence $\left|c_j^{(m)}\right| \leq C$ for all $j, m \in \mathbb{N}$. On the other hand we have

$$0 < \operatorname{Var}(Z_m) = \operatorname{Cov}\left(Z_m, \sum_{j=1}^{n} c_j^{(m)} Z_j\right) \leq C \sum_{j=1}^{n} |\mathbb{E}(Z_m Z_j)| \underset{m\to\infty}{\longrightarrow} 0.$$

□

We can now establish the asymptotic results for the AR parameter $\boldsymbol{\alpha}$ and the remaining terms. Throughout the proofs, the matrices $\mathbf{X}_0, \mathbf{W}_0, \mathbf{M}_0$ and $\widetilde{\mathbf{B}}_0$ are defined as $\mathbf{X}, \mathbf{W}, \mathbf{M}$ and $\widetilde{\mathbf{B}}$ but with kernel $K_0(\cdot)$ and bandwidth h_0. For notational convenience, with B_x and V_x we refer to $\mathbf{e}_1^\mathsf{T}\mathbf{B}_x$ and $\mathbf{e}_1^\mathsf{T}\mathbf{V}_x$ as defined in Chapter 2.

We first give the asymptotic order of the filtering parameter $\boldsymbol{\alpha}$.

Lemma 3.9 (Upper bound for least squares estimation of $\boldsymbol{\alpha}$)**.** *Under Assumptions 1 and 2, it holds* $\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} = O_P\left(h_0^{p+1} + \frac{1}{nh_0} + \frac{1}{\sqrt{n}}\right)$.

Proof. We decompose $\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}$ as follows:

$$\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} = \widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}} + \widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha},$$

where $\widetilde{\boldsymbol{\alpha}}$ denotes the least squares estimate of $\boldsymbol{\alpha}$ on the error terms $(U_t)_{t=1-q,\dots,n}$:

$$\widetilde{\boldsymbol{\alpha}} := \left(\mathbf{U}_q^\mathsf{T}\mathbf{U}_q\right)^{-1} \mathbf{U}_q^\mathsf{T}\mathbf{U}.$$

We prove that $\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha} = O_P\left(n^{-1/2}\right)$, and $\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}} = O_P\left(h_0^{p+1} + \frac{1}{nh_0}\right)$, which yields the assertion.

Asymptotic order of $\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}$. Since $\mathbf{U} = \mathbf{U}_q\boldsymbol{\alpha} + \mathbf{V}$ by definition, we get

$$\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha} = \left(\frac{1}{n}\mathbf{U}_q^{\mathsf{T}}\mathbf{U}_q\right)^{-1}\frac{1}{n}\mathbf{U}_q^{\mathsf{T}}\mathbf{V} =: \left(\mathbf{G}^{-1}\right)\frac{1}{n}\mathbf{U}_q^{\mathsf{T}}\mathbf{V}.$$

We first show that $\mathbf{G}^{-1} = O_P(1)$. Define $\boldsymbol{\Lambda} := \frac{1}{n}\mathbb{E}(\mathbf{U}_q^{\mathsf{T}}\mathbf{U}_q)$. Stationarity and the covariance inequality for strongly mixing processes (Lemma 2.5) provide

$$\begin{aligned}
\mathbb{E}([\mathbf{G}]_{k,l} - [\boldsymbol{\Lambda}]_{k,l})^2 &= \operatorname{Var}([\mathbf{G}]_{k,l}) \\
&= \operatorname{Var}\left(\frac{1}{n}\sum_{t=1}^{n} U_{t-k}U_{t-l}\right) \\
&\leq \frac{2}{n^2}\sum_{t\leq s}\left|\operatorname{Cov}\left(U_{t-k}U_{t-l}, U_{s-k}U_{s-l}\right)\right| \\
&\leq \frac{8}{n^2}\sum_{t\leq s}(\widetilde{\alpha_U}(s-t))^{1/3}\|U_{t-k}U_{t-l}\|_3\|U_{s-k}U_{s-l}\|_3 \\
&= O\left(\frac{1}{n}\right),
\end{aligned}$$

where

$$\widetilde{\alpha_U}(s-t) := \begin{cases} \alpha_U(s-t-|k-l|) & \text{if } s-t > |k-l|, \\ \alpha_U(0) & \text{otherwise,} \end{cases}$$

since by Assumption 2 (ii) the mixing coefficients $\alpha_U(z)^{1/3}$ are summable, and the sixth moments of $(U_t)_{t\in\mathbb{Z}}$ are finite. Hence by Chebyshev's inequality

$$\mathbf{G} - \boldsymbol{\Lambda} = O_P(n^{-1/2}). \tag{3.4}$$

Stationarity provides $[\boldsymbol{\Lambda}]_{k,l=1,\ldots,q} = \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}(U_{t-k}U_{t-l}) = \mathbb{E}(U_1U_{k-l+1})$, and thus

$$\boldsymbol{\Lambda} = \begin{pmatrix} \sigma_U^2 & \ldots & \mathbb{E}(U_1U_q) \\ \vdots & \ddots & \vdots \\ \mathbb{E}(U_1U_q) & \ldots & \sigma_U^2 \end{pmatrix}.$$

Again by virtue of Assumption 2 (ii) we get $|\mathbb{E}(U_1U_n)| \leq C(\alpha_U(n-1))^{1/3} \underset{n\to\infty}{\longrightarrow} 0$. Since $\sigma_U^2 > 0$, the matrix $\boldsymbol{\Lambda}$ is regular by virtue of Lemma 3.8. Equation (3.4) thus implies $\mathrm{P}($"$\mathbf{G}$ is positive definite"$) \underset{n\to\infty}{\longrightarrow} 1$. Since the matrix inverse is a continuous mapping on the set of regular matrices, the continuous mapping theorem gives $\mathbf{G}^{-1} - \boldsymbol{\Lambda}^{-1} = o_P(1)$, and consequently

$$\mathbf{G}^{-1} = O_P(1). \tag{3.5}$$

For the remaining term, we have for $i \in \{1, \dots, q\}$ arbitrary

$$\frac{1}{n}\left[\mathbf{U}_q^{\mathsf{T}}\mathbf{V}\right]_i = \frac{1}{n}\left[\mathbf{U}_q^{\mathsf{T}}(\mathbf{U} - \mathbf{U}_q\boldsymbol{\alpha})\right]_i = \frac{1}{n}\sum_{t=1}^{n} U_{t-i}(U_t - P_q U_t) =: S_n.$$

We show that $\mathbb{E}\left(S_n^2\right) = O\left(n^{-1}\right)$ which yields the desired order by Markov's inequality. Note that since $U_{t-i} \in \{U_{t-1}, \dots, U_{t-q}\}$, and since$P_q U_t$ is the orthogonal projection of U_t on the linear space spanned by $U_{t-1}, \dots, U_{t-q}$, we have $\mathbb{E}\left(U_{t-i}(U_t - P_q U_t)\right) = 0$, and consequently

$$\mathbb{E}\left(S_n^2\right) \leq \frac{2}{n^2}\sum_{t\leq s}\left|\mathrm{Cov}\left(U_{t-i}(U_t - P_q U_t), U_{s-i}(U_s - P_q U_s)\right)\right|.$$

The strong mixing property of $(U_t)_{t\in\mathbb{Z}}$ provides that the process $(U_{z-i}(U_z - P_q U_z))_{z\geq 1}$ is also strongly mixing with mixing coefficients

$$\widetilde{\widetilde{\alpha_U}}(k) := \begin{cases} \alpha_U(k-q), & \text{if } k \geq q, \\ \alpha_U(0) & \text{if } k < q. \end{cases}$$

Consequently, the covariance inequality for strongly mixing processes (Lemma 2.5) and stationarity provide

$$\begin{aligned}
&\left|\mathrm{Cov}\left(U_{t-i}(U_t - P_q U_t), U_{s-i}(U_s - P_q U_s)\right)\right| \\
&\leq 4\left(\widetilde{\widetilde{\alpha_U}}(s-t)\right)^{1/3}\|U_{t-i}(U_t - P_q U_t)\|_3^2 \\
&\leq C\left(\widetilde{\widetilde{\alpha_U}}(s-t)\right)^{1/3},
\end{aligned}$$

again by Hoelder's inequality and the finiteness of the sixth moments of $(U_t)_{t\in\mathbb{Z}}$. We thus conclude

$$\mathbb{E}\left(S_n^2\right) \leq \frac{2}{n^2}\sum_{t\leq s} C\,\left(\widetilde{\widetilde{\alpha_U}}(s-t)\right)^{1/3} = O\left(\frac{1}{n}\right),$$

again by the summability assumption on $(\alpha_U(\cdot))^{1/3}$. Applying Markov's inequality we obtain

$$\frac{1}{n}\mathbf{U}_q^{\mathsf{T}}\mathbf{V} = O_P\left(n^{-1/2}\right). \tag{3.6}$$

Equations (3.5) and (3.6) finally yield

$$\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha} = O_P\left(n^{-1/2}\right).$$

Asymptotic order of $\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}$. We decompose $\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}$ as follows:

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}} &= \widehat{\mathbf{G}}^{-1}\widehat{\mathbf{g}} - \mathbf{G}^{-1}\mathbf{g} \\
&= \mathbf{G}^{-1}(\widehat{\mathbf{g}} - \mathbf{g}) + (\widehat{\mathbf{G}}^{-1} - \mathbf{G}^{-1})\widehat{\mathbf{g}} \\
&= \mathbf{G}^{-1}(\widehat{\mathbf{g}} - \mathbf{g}) + (-\widehat{\mathbf{G}}^{-1}(\widehat{\mathbf{G}} - \mathbf{G})\mathbf{G}^{-1})(\mathbf{g} + \widehat{\mathbf{g}} - \mathbf{g}) \\
&= \mathbf{G}^{-1}(\widehat{\mathbf{g}} - \mathbf{g}) - \widehat{\mathbf{G}}^{-1}(\widehat{\mathbf{G}} - \mathbf{G})\mathbf{G}^{-1}\mathbf{g} - \widehat{\mathbf{G}}^{-1}(\widehat{\mathbf{G}} - \mathbf{G})\mathbf{G}^{-1}(\widehat{\mathbf{g}} - \mathbf{g}),
\end{aligned} \tag{3.7}$$

with $\mathbf{G}$ as defined before, and

$$\widehat{\mathbf{G}} := \frac{1}{n}\widehat{\mathbf{U}}_q^\mathsf{T}\widehat{\mathbf{U}}_q, \text{ and } \mathbf{g} := \frac{1}{n}\mathbf{U}_q^\mathsf{T}\mathbf{U}, \ \widehat{\mathbf{g}} := \frac{1}{n}\widehat{\mathbf{U}}_q^\mathsf{T}\widehat{\mathbf{U}}.$$

We show that both

$$\widehat{\mathbf{G}} - \mathbf{G} = O_P\left(h_0^{p+1}\right) + O_P\left(\frac{1}{nh_0}\right), \text{ and } \widehat{\mathbf{g}} - \mathbf{g} = O_P\left(h_0^{p+1}\right) + O_P\left(\frac{1}{nh_0}\right). \tag{3.8}$$

Equation (3.8) yields the desired convergence rate: we have $\widehat{\mathbf{G}} - \boldsymbol{\Lambda} = (\widehat{\mathbf{G}} - \mathbf{G}) + (\mathbf{G} - \boldsymbol{\Lambda})$, with $\boldsymbol{\Lambda}$ as defined in the last step. By equations (3.4) and (3.8), we thus get $\widehat{\mathbf{G}} - \boldsymbol{\Lambda} = o_P(1)$. Another usage of the continuous mapping theorem provides $\widehat{\mathbf{G}}^{-1} - \boldsymbol{\Lambda}^{-1} = o_P(1)$. Hence

$$\widehat{\mathbf{G}}^{-1} = O_P(1).$$

Similar calculations as for $\mathbf{G}$ give $\mathbf{g} = O_P(1)$. Therefore by equations (3.5), (3.7), and (3.8)

$$\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}} = O_P\left(h_0^{p+1}\right) + O_P\left(\frac{1}{nh_0}\right).$$

It remains to calculate the order of $\widehat{\mathbf{G}} - \mathbf{G}$ and $\widehat{\mathbf{g}} - \mathbf{g}$. We proceed element-wise. It holds $[\widehat{\mathbf{G}}]_{k,l} = \frac{1}{n}\sum_{t=1}^{n}\widehat{U}_{t-k}\widehat{U}_{t-l}$, and

$$\widehat{U}_t = Y_t - \widehat{m}(X_t) = U_t + (m(X_t) - \widehat{m}(X_t)) = U_t - B_{X_t} - V_{X_t}.$$

Therefore

$$\begin{aligned}
&\widehat{U}_{t-k}\widehat{U}_{t-l} - U_{t-k}U_{t-l} \\
&= (U_{t-k} - B_{X_{t-k}} - V_{X_{t-k}})(U_{t-l} - B_{X_{t-l}} - V_{X_{t-l}}) - U_{t-k}U_{t-l} \\
&= -U_{t-k}B_{X_{t-l}} - U_{t-l}B_{X_{t-k}} - U_{t-k}V_{X_{t-l}} - U_{t-l}V_{X_{t-k}} \\
&\quad + B_{X_{t-l}}B_{X_{t-k}} + V_{X_{t-l}}V_{X_{t-k}} + B_{X_{t-l}}V_{X_{t-k}} + V_{X_{t-l}}B_{X_{t-k}}.
\end{aligned}$$

We derive the following convergence rates for the mixed terms

$$\frac{1}{n}\sum_{t=1}^{n} B_{X_{t-l}}U_{t-k} = O_P\left(h_0^{p+1}\right), \quad \frac{1}{n}\sum_{t=1}^{n} V_{X_{t-l}}U_{t-k} = O_P\left(\frac{1}{n\sqrt{h_0}}\right), \tag{3.9}$$

and for the bias- and variance-type terms

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n} B_{X_{t-k}}B_{X_{t-l}} = O_P\left(h_0^{2(p+1)}\right), \quad \frac{1}{n}\sum_{t=1}^{n} B_{X_{t-k}}V_{X_{t-l}} = O_P\left(h_0^{p+1}\right), \\
&\frac{1}{n}\sum_{t=1}^{n} V_{X_{t-k}}V_{X_{t-l}} = O_P\left(\frac{1}{nh_0}\right).
\end{aligned} \tag{3.10}$$

With these convergence results we conclude

$$\begin{aligned}
|\widehat{\mathbf{G}}_{k,l} - \mathbf{G}_{k,l}| \leq{}& \left|\frac{1}{n}\sum_{t=1}^{n} U_{t-k}B_{X_{t-l}} + U_{t-l}B_{X_{t-k}}\right| + \left|\frac{1}{n}\sum_{t=1}^{n} U_{t-k}V_{X_{t-l}} + U_{t-l}V_{X_{t-k}}\right| \\
&+ \left|\frac{1}{n}\sum_{t=1}^{n}\left(B_{X_{t-l}}B_{X_{t-k}} + B_{X_{t-l}}V_{X_{t-k}} + V_{X_{t-l}}B_{X_{t-k}} + V_{X_{t-l}}V_{X_{t-k}}\right)\right| \\
={}& O_P\left(h_0^{p+1}\right) + O_P\left(\frac{1}{nh_0}\right).
\end{aligned}$$

The calculations for $\mathbf{g}$ are similar, and hence the proof is completed.

We are left to prove equations (3.9) and (3.10). The uniform convergence results for the bias term provided by Corollary 2.16, namely

$$\sup_{x\in D} B_x = O_P\left(h_0^{p+1}\right),$$

imply the asymptotic order for those terms that include the bias. The remaining ("variance"-) terms require a more detailed analysis.

We first take a look at the mixed terms in equation (3.9):

Asymptotic order of first mixed term: $\frac{1}{n}\sum_{t=1}^{n} B_{X_{t-l}}U_{t-k} = O_P\left(h_0^{p+1}\right)$. The uniform convergence results for the bias term and the boundedness of the second moments of $(U_t)_{t\in\mathbb{Z}}$ directly imply $\mathbb{E}\left|\frac{1}{n}\sum_{t=1}^{n} B_{X_{t-l}}U_{t-k}\right| = O\left(h_0^{p+1}\right)$. The assertion now follows from Markov's inequality.

Asymptotic order of second mixed term: $\frac{1}{n}\sum_{t=1}^{n} V_{X_{t-l}}U_{t-k} = O_P\left(\frac{1}{n\sqrt{h_0}}\right)$. By virtue of Corollary 2.14 and Assumption 1 (i) on the strong mixing coefficients $\alpha_X(\cdot)$ and the density

$f_X(\cdot)$ we have $(\mathbf{X}_0^T\mathbf{W}_0\mathbf{X}_0)^{-1} \xrightarrow{a.s.} \mathbf{M}_0{}^{-1}f_X(x)^{-1}$ uniformly in $x \in D$. Therefore

$$\begin{aligned}
V_x &= \mathbf{e}_1^\mathsf{T}(\mathbf{X}_0^T\mathbf{W}_0\mathbf{X}_0)^{-1}\mathbf{X}_0^T\mathbf{W}_0\mathbf{U} \\
&= \mathbf{e}_1^\mathsf{T}\,\mathbf{M}_0{}^{-1}f_X(x)^{-1}\mathbf{X}_0^T\mathbf{W}_0\mathbf{U}(1+o_P(1)) \\
&= (1+o_P(1))\mathbf{e}_1^\mathsf{T}\,\mathbf{M}_0{}^{-1}f_X(x)^{-1}\begin{pmatrix} \frac{1}{nh_0}\sum\limits_{i=1}^{n} K_0\left(\frac{X_i-x}{h_0}\right)U_i \\ \vdots \\ \frac{1}{nh_0}\sum\limits_{k=1}^{n} K_0\left(\frac{X_i-x}{h_0}\right)\left(\frac{X_i-x}{h_0}\right)^p U_i \end{pmatrix},
\end{aligned}$$

uniformly in $x \in D$.

Since $(\mathbf{M}_0)^{-1}$ is a finite $(p+1)\times(p+1)$ matrix and $f_X(\cdot)$ is bounded away from zero, it thus suffices to establish the asymptotic order of the following term:

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n}\left[\mathbf{X}_0^T\mathbf{W}_0\mathbf{U}\right]_r U_{t-k} &= \frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{n}\frac{1}{nh_0}K_0\left(\frac{X_i-X_{t-l}}{h_0}\right)\left(\frac{X_i-X_{t-l}}{h_0}\right)^{r-1}U_iU_{t-k} \\
&=: \frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{n}\frac{1}{nh_0}C_{K_0}(X_i,X_{t-l})\,U_iU_{t-k},
\end{aligned}$$

for some arbitrary $r \in \{1,\ldots,p+1\}$.

Since $(X_t)_{t\in\mathbb{Z}}$ and $(U_t)_{t\in\mathbb{Z}}$ are independent, we get

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{n}\frac{1}{nh_0}C_{K_0}(X_i,X_{t-l})U_iU_{t-k}\right)^2 \\
&= \frac{1}{n^2}\sum_{i,j,s,t}\frac{1}{n^2h_0^2}\mathbb{E}\left(C_{K_0}(X_i,X_{t-l})C_{K_0}(X_j,X_{s-l})\right)\mathbb{E}(U_iU_jU_{s-k}U_{t-k}).
\end{aligned}$$

Hoelder's inequality and stationarity provide

$$\mathbb{E}\left(C_{K_0}(X_i,X_{t-l})C_{K_0}(X_j,X_{s-l})\right) \le \mathbb{E}\left(C_{K_0}^2(X_i,X_{t-l})\right) = O\left(h_0\right),$$

since for all i,t we have

$$\begin{aligned}
\mathbb{E}\left(C_{K_0}^2(X_i,X_{t-l})\right) &= \int_{u,v}K_0^2\left(\frac{u-v}{h_0}\right)\left(\frac{u-v}{h_0}\right)^{2r-2}f_{X_i,X_{t-l}}(u,v)\,du\,dv \\
&= \int_v\left(\int_z K_0^2\left(z\right)z^{2r-2}f_{X_i|X_{t-l}}(v+h_0z)h_0dz\right)f_{X_{t-l}}(v)\,dv \\
&= O\left(h_0\right),
\end{aligned}$$

since by virtue of Assumptions 1 and 2 the product and conditional densities of the design process are bounded and the Kernels are bounded with compact support.

Further, it holds

$$\begin{aligned}\mathbb{E}(U_iU_jU_{s-k}U_{t-k}) = \mathrm{Cum}(U_i,U_j,U_{s-k},U_{t-k}) + \mathbb{E}(U_iU_j)\mathbb{E}(U_{t-k}U_{s-k})\\ + \mathbb{E}(U_iU_{t-k})\mathbb{E}(U_jU_{s-k}) + \mathbb{E}(U_iU_{s-k})\mathbb{E}(U_jU_{t-k}).\end{aligned} \tag{3.11}$$

Lemma 3.7 yields that we have with $v := s-k$ and $w := t-k$

$$\begin{aligned}&\sum_{i,j,v,w} |\mathrm{Cum}(U_i,U_j,U_v,U_w)|\\ &\leq 4! \sum_{1\leq i\leq j\leq v\leq w\leq n} |\mathrm{Cum}(U_i,U_j,U_v,U_w)|\\ &\leq 4! \sum_{1\leq i\leq j\leq v\leq w\leq n} C\ \alpha_U(\max\{j-i,v-j,w-v\})^{1/3}\\ &= 4! \sum_{T=0}^{n} C\ \alpha_U(T)^{1/3} \#\{(1\leq i\leq j\leq v\leq w\leq n) : \max\{j-i,v-j,w-v\}=T\}\\ &\leq C\ n \sum_{T=0}^{n} \alpha_U(T)^{1/3}(T+1)^2\\ &= O(n),\end{aligned}$$

by the summability condition on the strong mixing coefficients $\alpha_U(\cdot)$. We used that the number of time points $(1\leq i\leq j\leq v\leq w\leq n)$ with maximal distance T can be calculated as follows:

$$\begin{aligned}&\#\{(1\leq i\leq j\leq v\leq w\leq n) : \max\{j-i,v-j,w-v\}=T\}\\ &= 3\times\#\{i-j=T \text{ and } v\in\{j,\ldots,j+T\}, w\in\{v,\ldots,v+T\}\}\\ &\leq 3\,n(T+1)^2,\end{aligned}$$

since there are three possibilities where the maximum distance T can be taken.

For the remaining terms, Lemma 3.6 provides

$$\sum_{1\leq t_1,t_2,t_3,t_4\leq n} \mathbb{E}(U_{t_1}U_{t_2})\mathbb{E}(U_{t_3}U_{t_4}) = \left(\sum_{1\leq t_1,t_2\leq n} \mathbb{E}(U_{t_1}U_{t_2})\right)^2 = O\left(n^2\right).$$

And thus by equation (3.11)

$$\mathbb{E}\left(\frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{n}\frac{1}{nh_0}C_{K_0}(X_i,X_{t-l})U_iU_{t-k}\right)^2 = \frac{1}{n^2}O\left(n^2\right)O\left(\frac{1}{n^2h_0}\right) = O\left(\frac{1}{n^2h_0}\right).$$

The assertion now follows from Markov's inequality.

Now we take a closer look at the terms in equation (3.10).

Asymptotic order of bias-type term: $\boldsymbol{\frac{1}{n}\sum_{t=1}^{n} B_{X_{t-k}} B_{X_{t-l}} = O_P\left(h_0^{2(p+1)}\right)}$**.** This is a direct consequence of the uniform convergence result for B_x.

Asymptotic order of bias-variance-type term: $\boldsymbol{\frac{1}{n}\sum_{t=1}^{n} B_{X_{t-k}} V_{X_{t-l}} = O_P\left(h_0^{p+1}\right)}$**.** The uniform convergence results for the bias term and Markov's inequality give

$$\left|\frac{1}{n}\sum_{t=1}^{n} B_{X_{t-k}} V_{X_{t-l}}\right| \leq \sup_{x\in D} |B_x| \frac{1}{n}\sum_{t=1}^{n} |V_{X_{t-k}}| = O_P\left(h_0^{p+1}\right) O_P(1),$$

since in analogy to previous calculations we have that $V_{X_{t-k}} = O_P(1)\mathbf{X}_0^T \mathbf{W}_0 \mathbf{U} = O_P(1)$ by a direct calculation of expectations.

Asymptotic order of variance-type term: $\boldsymbol{\frac{1}{n}\sum_{t=1}^{n} V_{X_{t-k}} V_{X_{t-l}} = O_P\left(\frac{1}{nh_0}\right)}$**.** In complete analogy to the calculations for the second mixed term and with the same arguments as stated therein, it suffices to study the order of the following term:

$$\frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \omega_{t,i,k,m}\, \omega_{t,j,l,o}\, U_i U_j,$$

with $\omega_{t,i,k,m} := \frac{1}{nh_0} K_0\left(\frac{X_i - X_{t-k}}{h_0}\right)\left(\frac{X_i - X_{t-k}}{h_0}\right)^{m-1}$, and $l, m \in \{1, \ldots, p+1\}$ arbitrary.

Hölder's inequality provides

$$\mathbb{E}\left|\sum_{i=1}^{n}\sum_{j=1}^{n} \omega_{t,i,k,m}\omega_{t,j,l,o}\, U_i U_j\right| \leq \sqrt{\mathbb{E}\left(\left(\sum_{i=1}^{n} \omega_{t,i,k,m}\, U_i\right)^2\right) \mathbb{E}\left(\left(\sum_{j=1}^{n} \omega_{t,j,l,o}\, U_j\right)^2\right)}.$$

Since $(X_t)_{t\in\mathbb{Z}}$ and $(U_t)_{t\in\mathbb{Z}}$ are independent, we get

$$\mathbb{E}\left(\left(\sum_{i=1}^{n} \omega_{t,i,k,m}\, U_i\right)^2\right) = \sum_{i,r} \mathbb{E}\left(\omega_{t,i,k,m}\, \omega_{t,r,k,m}\right) \mathbb{E}(U_i U_r) = O\left(\frac{1}{nh_0}\right),$$

since $\mathbb{E}\left(\omega_{t,i,k,m}\, \omega_{t,r,k,m}\right) = O\left(\frac{1}{n^2 h_0}\right)$ similar to former calculations, and $\sum_{i,r} \mathbb{E}(U_i U_r) = O(n)$ by Lemma 3.6.

Thus

$$\mathbb{E}\left|\frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \omega_{t,i,k,m}\, \omega_{t,j,l,o}\, U_i U_j\right| = O\left(\frac{1}{nh_0}\right),$$

and hence by Markov's inequality $\frac{1}{n}\sum_{t=1}^{n} V_{X_{t-k}} V_{X_{t-l}} = O_P\left(\frac{1}{nh_0}\right)$. □

We can now provide asymptotic orders for the remaining terms.

Lemma 3.10. *Under Assumptions 1 and 2, it holds* $\mathbf{R}_{n,1} = o_P\left(\frac{1}{nh}\right)$.

Proof. We have

$$|[\mathbf{R}_{n,1}]_r| \leq \max_k |\widehat{\alpha}_k - \alpha_k| \sum_{k=1}^{q} \left| \frac{1}{nh} \sum_{t=1}^{n} C_K(X_t) U_{t-k} \right|,$$

with $C_K(y) := K\left(\frac{y-x}{h}\right)\left(\frac{y-x}{h}\right)^{r-1}$.

Lemma 3.9 and Assumption 2 (iv) on the bandwidths provide $\max_k |\widehat{\alpha}_k - \alpha_k| = o_P\left(\frac{1}{\sqrt{nh}}\right)$. By verification of moments, we show that $\frac{1}{nh}\sum_{t=1}^{n} C_K(X_t)U_{t-k} = O_P\left(\frac{1}{\sqrt{nh}}\right)$, which yields the assertion.

We have $\mathbb{E}\left(\frac{1}{nh}\sum_{t=1}^{n} C_K(X_t)U_{t-k}\right) = 0$, and since the processes $(X_t)_{t\in\mathbb{Z}}$ and $(U_t)_{t\in\mathbb{Z}}$ are both stationary and independent, it holds

$$\begin{aligned}
&\operatorname{Var}\left(\frac{1}{nh}\sum_{t=1}^{n} C_K(X_t)U_{t-k}\right) \\
&\leq \frac{1}{nh^2}\operatorname{Var}\left(C_K(X_1)U_{1-k}\right) + \frac{2}{nh^2}\sum_{t=2}^{n}\left|\operatorname{Cov}\left(C_K(X_1)U_{1-k}, C_K(X_t)U_{t-k}\right)\right| \\
&= \frac{1}{nh^2}\sigma_U^2 \mathbb{E}\left(C_K^2(X_1)\right) + \sum_{t=2}^{n} \left|\mathbb{E}\left(C_K(X_1)C_K(X_t)\right)\right| |\mathbb{E}(U_{1-k}U_{t-k})|.
\end{aligned}$$

Since the kernels are bounded and have compact support, and all product and conditional densities of the design process are bounded by Assumptions 1 (iii) and 2 (iii), respectively, we have $\mathbb{E}\left(C_K^2(X_1)\right) = O(h)$. Further, for $t > 1$

$$\mathbb{E}\left(C_K(X_1)C_K(X_t)\right) = h^2 \int K(u)u^{r-1}K(v)v^{r-1} f_{X_1,X_t}(x+hu, x+hv)du = O(h^2).$$

Since $(U_t)_{t\in\mathbb{Z}}$ is alpha-mixing and centered, the covariance inequality for strongly mixing processes (Lemma 2.5) provides

$$|\mathbb{E}(U_{1-k}U_{t-k})| \leq 4\,\alpha_U(t-1)^{1/3} \sup_s \|U_s\|_3^2 \leq C\,\alpha_U(t-1)^{1/3},$$

since the third moments of the error process are finite by Assumption 2 (ii).

Hence

$$\operatorname{Var}\left(\frac{1}{nh}\sum_{t=1}^{n} C_K(X_t)U_{t-k}\right) = O\left(\frac{1}{nh}\right) + O(1)\,\frac{1}{n}\sum_t \alpha_U(t-1)^{1/3} = O\left(\frac{1}{nh}\right),$$

since by Assumption 2 (ii) the mixing coefficients $(\alpha_U(\cdot))^{1/3}$of the error process are summable. Chebyshev's inequality thus provides $\frac{1}{nh}\sum\limits_{t=1}^{n} C_K(X_t)U_{t-k} = O_P\left(\frac{1}{\sqrt{nh}}\right)$. □

Lemma 3.11. *Under Assumptions 1 and 2, it holds* $\mathbf{R}_{n,2} = O_P(h_0^{p+1}) + o_P\left(\frac{1}{\sqrt{nh}}\right)$.

Proof. We decompose $[\mathbf{R}_{n,2}]_r$ in a bias term

$$\mathbf{R}^B_{n,2,r} := \frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1}\sum_{k=1}^{q}\alpha_k B_{X_{t-k}},$$

and a variance term

$$\mathbf{R}^V_{n,2,r} := \frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1}\sum_{k=1}^{q}\alpha_k V_{X_{t-k}}.$$

The proof is completed by showing that $\mathbf{R}^B_{n,2,r} = O_P(h_0^{p+1})$ and $\mathbf{R}^V_{n,2,r} = o_P\left(\frac{1}{\sqrt{nh}}\right)$.

For the bias term, Corollary 2.16 yields $\sup_{x\in D} B_x = O_P\left(h_0^{p+1}\right)$. Further, by Markov's inequality $\frac{1}{nh}\sum_{t=1}^{n}\left|K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^{r-1}\right| = O_P(1)$, which yields the asymptotic order of $\mathbf{R}^B_{n,2,r}$.

The variance term requires a more profound analysis. With the same convergence arguments for $V_{X_{t-k}}$ as used in Lemma 3.9 it suffices to study the order of

$$\begin{aligned} S_n &:= \frac{1}{n}\sum_{i=1}^{n} U_i \frac{1}{n}\sum_{t=1}^{n}\frac{1}{h_0 h}K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1} K_0\left(\frac{X_i - X_{t-k}}{h_0}\right)\left(\frac{X_i - X_{t-k}}{h_0}\right)^{s-1} \\ &=: \frac{1}{n}\sum_{i=1}^{n} U_i \frac{1}{n}\sum_{t=1}^{n}\omega_{t,i}, \end{aligned}$$

for an arbitrary $s \in \{1, \ldots, p+1\}$.

We now decompose the magnitude of S_n as follows:

$$\begin{aligned} \mathbb{E}\left(S_n^2\right) = &\frac{1}{n^4}\sum_{t,i}\mathbb{E}\left(\omega_{t,i}^2\right)\mathbb{E}(U_i^2) + \frac{1}{n^4}\sum_{t\neq s}\sum_{i}\mathbb{E}\left(\omega_{t,i}\omega_{s,i}\right)\mathbb{E}(U_i^2) \\ &+ \frac{1}{n^4}\sum_{t}\sum_{i\neq j}\mathbb{E}\left(\omega_{t,i}\omega_{t,j}\right)\mathbb{E}(U_iU_j) + \frac{1}{n^4}\sum_{t\neq s}\sum_{i\neq j}\mathbb{E}\left(\omega_{t,i}\omega_{s,j}\right)\mathbb{E}(U_iU_j). \end{aligned}$$

Since the error process has bounded second moments we have $\sum_{i=1}^{n}\mathbb{E}(U_i^2) = O(n^2)$. Further, Lemma 3.6 gives $\sum_{i\neq j}^{n}\mathbb{E}(U_iU_j) = O(n)$, Regarding the order of the kernel product, we have for $t \neq i$

$$
\begin{aligned}
&\mathbb{E}\left(\omega_{t,i}^2\right)\\
&= \frac{1}{h_0^2 h^2} \int K^2\left(\frac{u-x}{h}\right)\left(\frac{u-x}{h}\right)^{2r-2} K_0^2\left(\frac{v-w}{h_0}\right)\left(\frac{v-w}{h_0}\right)^{2s-2}\\
&\quad \times f_{X_t,X_i,X_{t-k}}(u,v,w)\, du\, dv\, dw\\
&= \frac{1}{h_0^2 h^2} \int K^2(y)\, y^{2r-2} K_0^2(z)\, z^{2s-2} f_{(X_t,X_i)|X_{t-k}=w}(x+hy, w+h_0 z)\, h_0\, h\, dy\, dz\\
&\quad \times \int f_{X_{t-k}}(w)\, dw\\
&= O\left(\frac{1}{h_0 h}\right),
\end{aligned}
$$

since by Assumption 1 (iii) all densities of the design process are bounded and the Kernels are bounded and have compact support. Similarly, $\mathbb{E}\left(\omega_{t,t}^2\right) = O\left(\frac{1}{h_0 h}\right)$.

If all indices $i, j, s-k, s, t-k, t$ are pairwise different, we obtain

$$
\begin{aligned}
&\left|\mathbb{E}\left(\omega_{t,i}\omega_{s,j}\right)\right|\\
&= \Bigg|\mathbb{E}\Bigg(\frac{1}{h_0^2 h^2} K_0\left(\frac{X_i - X_{t-k}}{h_0}\right)\left(\frac{X_i - X_{t-k}}{h_0}\right)^{s-1} K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1}\\
&\quad \times K_0\left(\frac{X_j - X_{s-k}}{h_0}\right)\left(\frac{X_j - X_{s-k}}{h_0}\right)^{s-1} K\left(\frac{X_s - x}{h}\right)\left(\frac{X_s - x}{h}\right)^{r-1}\Bigg)\Bigg|\\
&\leq \frac{1}{h_0^2 h^2} C_f \int_{u,v,w,y} \left|K_0(u)\, u^{s-1} K(v)\, v^{r-1} K_0(w)\, w^{s-1} K(y)\, y^{r-1}\right| h_0^2\, h^2 du\, dv\, dw\, dy\\
&\quad \times \int_{z,s} f_{X_i,X_j}(z,s)\, dz\, ds\\
&= O(1),
\end{aligned}
$$

where C_f is a constant bounding the conditional density $f_{(X_{s-k},X_{t-k},X_s,X_t)|(X_i,X_j)=(z,s)}(\cdot)$. With similar calculations we obtain $\mathbb{E}\left(\omega_{t,i}\omega_{s,j}\right) = O(1)$ for all remaining cases in which $i \neq j$ and $s \neq t$. For $t \neq s$, but $i = j$, we get in complete analogy $\mathbb{E}\left(\omega_{t,i}\omega_{s,i}\right) = O(1)$, and for $t = s$ but $i \neq j$, we get $\mathbb{E}\left(\omega_{t,i}\omega_{t,j}\right) = O\left(\frac{1}{h}\right)$.

We thus obtain

$$
\mathbb{E}\left(\frac{1}{n^2}\sum_{t,i=1}^{n} \omega_{t,i} U_i\right)^2 = O\left(\frac{1}{n^2 h_0 h} + \frac{1}{n} + \frac{1}{n^2 h} + \frac{1}{n}\right) = O\left(\frac{1}{n^2 h_0 h} + \frac{1}{n}\right),
$$

and hence applying Markov's inequality

$$
\mathbf{R}_{n,2,r}^V = O_P\left(\frac{1}{n\sqrt{h_0 h}}\right) + O_P\left(\frac{1}{\sqrt{n}}\right) = o_P\left(\frac{1}{\sqrt{nh}}\right).
$$

□

Lemma 3.12. *Under Assumptions 1 and 2, it holds* $\mathbf{R}_{n,3} = o_P\left(\frac{h_0^{p+1}}{\sqrt{nh}}\right) + o_P\left(\frac{1}{nh}\right)$.

Proof. It holds

$$|[\mathbf{R}_{n,3}]_r| \leq \sum_{k=1}^{q} |\alpha_k - \widehat{\alpha}_k| \left| \frac{1}{nh} \sum_{t=1}^{n} K\left(\frac{X_t - x}{h}\right) \left(\frac{X_t - x}{h}\right)^{r-1} (\widehat{m}(X_{t-k}) - m(X_{t-k})) \right|.$$

From Lemma 3.9 and Assumption 2 (iv) on the bandwidths, we get $\max_{k \geq 1} |\alpha_k - \widehat{\alpha}_k| = o_P\left(\frac{1}{\sqrt{nh}}\right)$, and thus the assertion follows from Lemma 3.11. □

Chapter 4

More efficient estimation in the heteroscedastic case

We extend the previous results to a heteroscedastic regression model. We show that even in this setting we can construct a more efficient estimator using the filtering technique provided in the previous chapter. In the first instance we introduce the estimation idea and a feasible estimation procedure. A prior estimation of the variance will be needed, for which we obtain uniform convergence results in the second part of this chapter. The convergence of the proposed estimator is shown in the last instance of this chapter.

Contents

4.1 Motivation

In what follows, we want to weaken the assumption of a constant error variance. To understand why this is useful, consider a simple example: imagine you would observe a car and measure the distance that it travels each second. If the car departs right next to you, your observations maybe correct to centimeters. However, as its distance to you increases, your measurements might only be good to some hundred meters. The data you collect would thus have a changing error variance.

When the error variance is not constant, we speak of heteroscedasticity. Heteroscedastic datasets are frequently encountered in nearly all fields, like econometrics (Greene 1993, Marno 2004), or ecological aspects including genetics (Brem & Kruglyak 2005, Daye & Chen 2012), toxicology (Lim et al. 2010), and fisheries research (Caroll & Ruppert 1988). Typically, a large disparity between the observed values causes the data to be heteroscedastic, since small values can usually be obtained with a smaller measurement error than large values.

A widely used regression model that considers heteroscedasticity is the following:

$$Y_t = m(X_t) + \sigma(X_t)U_t, \quad t = 1, \dots, n,$$

with $(X_t)_{t\in\mathbb{Z}}$ and $(U_t)_{t\in\mathbb{Z}}$ independent stationary processes, $\mathbb{E}(U_t) = 0$, $\mathrm{Var}(U_t) = 1$, and $0 < C_{\sigma,1} \leq \sigma(x) \leq C_{\sigma,2}$ for all $x \in \mathbb{R}$. In what follows, we develop a more efficient estimator $\widetilde{m}^{\text{het}}(x)$ for this regression model by adapting the filtering technique introduced in the previous chapter to this regression model. In what follows, we show that our proposed estimator satisfies

$$\sqrt{nh}\left(\widetilde{m}^{\text{het}}(x) - m(x) - h_n^{p+1}B(x)\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma^2_{\overline{U}^{\text{het}}}\sigma^2(x)V(x)\right),$$

with $\sigma^2_{\overline{U}^{\text{het}}} \leq 1$. Our estimator has thus an asymptotically smaller variance as the conventional estimator but the same bias and is consequently more efficient.

To the best of the authors' knowledge, no related research on a more-stage method for this heteroscedastic regression model can be found in the literature. However, Su et al. (2012) proposed a two-stage method for a heteroscedastic linear regression model including local polynomial regression to estimate the variance. In a simulation study, their estimator provided more precise estimations than the standard OLS and GLS. Our proposed estimation procedure can be seen as a natural extension of their procedure to the nonparametric regression model.

4.2 Estimation Method in the heteroscedastic case

The estimation idea introduced in the previous chapter can easily be carried over to the heteroscedastic setting. As in the homoscedastic case, we obtain a more efficient estimator by filtering the error process, for which we consider a linear approximation $P_qU_t := \alpha_1 U_{t-1} + \ldots + \alpha_q U_{t-q}$ of U_t. The filtering parameters $\alpha_1, \dots, \alpha_q$ are again chosen to minimize the mean squared error $\mathbb{E}(U_t - P_qU_t)^2$.

Define the filtered series

$$\overline{Y_t}^{\text{het}} := Y_t - \sigma(X_t)P_qU_t,$$

and $\widehat{m}(x)$ and $\overline{m}^{\text{het}}(x)$ the pth order local polynomial estimators of $m(x)$ based on Y_t and $\overline{Y_t}^{\text{het}}$, respectively.

Corollary 2.23 provides the convergence of both estimators

$$\boxed{\begin{array}{c}\sqrt{nh_n}\left(\widehat{m}(x) - m(x) - h_n^{p+1}B(x)\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma_U^2\sigma^2(x)V(x)\right), \\ \sqrt{nh_n}\left(\overline{m}^{\text{het}}(x) - m(x) - h_n^{p+1}B(x)\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma^2_{\overline{U}^{\text{het}}}\sigma^2(x)V(x)\right),\end{array}}$$

with the bias term $B(x)$ and $V(x)$ as in the homoscedastic case, the filtered error $\overline{U}_t^{\text{het}} := U_t - P_qU_t$, and its variance function $\sigma^2_{\overline{U}^{\text{het}}} := \mathbb{E}\left(\overline{U}_t^{\text{het}}\right)^2$.

As in the homoscedastic case, it holds for the variances of the error processes

$$\begin{aligned}1 &= \mathbb{E}\, U_t^2 \\ &= \mathbb{E}(U_t - P_qU_t)^2 + \mathbb{E}(P_qU_t)^2 + 2\,\mathbb{E}(U_t - P_qU_t)(P_qU_t) \\ &\geq \sigma^2_{\overline{U}^{\text{het}}}.\end{aligned}$$

For $\mathrm{P}(P_qU_t \neq 0) > 0$, we have $\mathbb{E}(P_qU_t)^2 > 0$, and consequently

$$\sigma^2_{\overline{U}^{\text{het}}} < 1.$$

We conclude that also in the heteroscedastic setting the efficiency of the estimator $\widehat{m}(x)$ can be improved by using $\overline{m}^{\text{het}}(x)$ instead.

Similar to the homoscedastic case, the error process $(U_t)_{t\in\mathbb{Z}}$ and the parameter vector $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_q)^{\mathsf{T}}$ are unknown. In this regression model we additionally have to estimate the unknown conditional variance $\sigma^2(x)$. We will use a residual based approach for this purpose, which means that we estimate $\sigma^2(x)$ via local polynomial smoothing on the squared residuals $(Y_t - \widehat{m}(X_t))^2$ that we obtain in a prior smoothing. In Section 4.3, we explain the choice of this estimation approach in detail and study the asymptotic behavior of the resulting estimator $\widehat{\sigma^2}(x)$.

The resulting estimation algorithm is a three-step procedure, since we execute three local polynomial fits: first the prior smoothing, second the smoothing on the squared residuals, and third the final fit. The estimation procedure can be formulated as follows:

Estimation Procedure (Proposed heteroscedastic estimator)**.**

1. *Pilot fit:* Obtain a preliminary local pth order polynomial smoothing Y_t on X_t, using kernel $K_0(\cdot)$ and bandwidth $h_0 = h_0(n)$. Denote the estimates $\widehat{m}(X_t)$ and calculate an approximation of the errors $\sigma(X_t)U_t$ by the residuals

$$\widehat{\sigma(X_t)U_t} = Y_t - \widehat{m}(X_t), \qquad t = 1, \ldots, n.$$

2. *Approximation of variance function:* Obtain an estimate $\widehat{\sigma^2}(X_t)$ of the variance function $\sigma^2(x)$ in $x = X_t$ for $t = 1, \ldots, n$ by local pth order polynomial smoothing of $(\widehat{\sigma(X_t)U_t})^2$

on X_t, using kernel $K_\sigma(\cdot)$ and bandwidth $h_\sigma = h_\sigma(n)$. Set $\widehat{\sigma}(X_t) := \sqrt{\max\left\{0, \widehat{\sigma^2}(X_t)\right\}}$.

3. *Calculation of filter parameters:* Approximate the error process as follows:

$$\widehat{U_t} = \frac{\widehat{\sigma(X_t)U_t}}{\widehat{\sigma}(X_t)}, \qquad t = 1, \ldots, n.$$

   Calculate an estimate $\widehat{\boldsymbol{\alpha}}$ of the filter parameters $\boldsymbol{\alpha}$ by least-squares estimation on $\widehat{U_t}$, $t = 1, \ldots, n$ as conducted in the homoscedastic case.

4. *Prewhitening:* Calculate an approximation $\widetilde{\mathbf{Y}}^{\text{het}} = (\widetilde{Y}_1^{\text{het}}, \ldots, \widetilde{Y}_n^{\text{het}})$ of the filtered series $\overline{\mathbf{Y}}^{\text{het}}$ by

$$\widetilde{Y}_t^{\text{het}} = Y_t - \widehat{\sigma}(X_t)\left(\widehat{\alpha}_1\widehat{U}_{t-1} - \ldots - \widehat{\alpha}_q\widehat{U}_{t-q}\right), \qquad t = 1, \ldots, n.$$

5. *Final fit:* The proposed estimator $\widetilde{m}^{\text{het}}(x)$ is then obtained by pth order local polynomial smoothing of $\widetilde{Y}_t^{\text{het}}$ on X_t, using kernel $K(\cdot)$ and bandwidth $h = h(n)$.

The foregoing procedure may be iterated to achieve better finite-sample performance in practice. In Section 4.4 we prove that the feasible estimator $\widetilde{m}^{\text{het}}(x)$ is asymptotically equivalent to the infeasible estimator $\overline{m}^{\text{het}}(x)$, which we have already shown to be more efficient than the conventional estimator $\widehat{m}(x)$. It should thus be preferred.

Throughout this chapter, we suppose Assumption 1 to hold. Additionally, we make the following assumptions:

Assumption 3.

(i) The variance function $\sigma^2(\cdot)$ is $(p+1)$-times differentiable and the derivatives are continuous and bounded.

(ii) It holds $(X_t)_{t\in\mathbb{Z}} \in D$ with D a compact interval. The density $f_X(\cdot)$ of $(X_t)_{t\in\mathbb{Z}}$ is Lipschitz continuous.

(iii) The error process has finite ninth moments: $\|U_t\|_9 < \infty$.

(iv) The strong mixing coefficients $\alpha_X(\cdot)$ and $\alpha_U(\cdot)$ of $(X_t)_{t\in\mathbb{Z}}$ and $(U_t)_{t\in\mathbb{Z}}$ are exponentially decreasing, i.e. they satisfy $\alpha_X(k) \leq \exp(-C_X k)$, and $\alpha_U(k) \leq \exp(-C_U k)$.

(v) The kernel functions $K_0(\cdot)$, $K_\sigma(\cdot)$ and $K(\cdot)$ are nonnegative, Lipschitz continuous, bounded, strictly positive in zero, and have support on $[-1, 1]$.

(vi) The bandwidths h_0, h_σ and h, where $h_\sigma/h_0 \underset{n\to\infty}{\longrightarrow} C \in (0, \infty)$, are all $o(1)$, but satisfy $\sqrt{n}h_0 \underset{n\to\infty}{\longrightarrow} \infty$, and $nh \underset{n\to\infty}{\longrightarrow} \infty$. Further, it holds $\sqrt{nh}h_0^{p+1} \underset{n\to\infty}{\longrightarrow} 0$, and $\log(n)\sqrt{h}/(\sqrt{n}h_0) \underset{n\to\infty}{\longrightarrow} 0$.

Remark 4.1 (Some notes on the assumptions)**.** *Under Assumption 3 (iv) on the mixing coefficients, the serial dependence in the data decreases sufficiently fast to apply Bernstein's inequality for strongly mixing processes (Lemma 4.2). We will need this inequality in order to calculate uniform rates of convergence for both the conventional local polynomial estimator $\widehat{m}(\cdot)$ (see Lemma 4.3) and the variance estimator $\widehat{\sigma^2}(\cdot)$ (see Lemma 4.7). In both cases, we achieve the usual uniform convergence rates for local polynomial regression under mixing conditions, as stated for example in Masry (1996*a*).*

One important reason for our procedure to work is that the bias from both prior estimations can be made asymptotically negligible by undersmoothing in the first stage. This is guaranteed by Assumption 3 (vi) on the bandwidths ($\sqrt{nh}h_0^{p+1} \to 0$). The requirement $\sqrt{n}h_0 \to \infty$ is needed for the uniform convergence of both the conventional local polynomial estimator and the variance estimator, respectively. Among many possible bandwidth choices, we can select the main bandwidth h in a way to minimize the asymptotic mean squared error of the proposed estimator, namely

$$h \asymp n^{-\frac{1}{2p+3}}.$$

With this choice of h, we choose the prior bandwidths h_0 and h_σ of slightly smaller order than h

$$h_0 = C_{h_0} h^{\delta}, \text{and } h_\sigma = C_{h_\sigma} h^{\delta}, \text{ with } 1 < \delta < 2.5,$$

in order to satisfy all bandwidth conditions.[1]

To facilitate the notation, we assumed that the two prior bandwidths have the same order ($h_\sigma/h_0 \underset{n\to\infty}{\longrightarrow} C \in (0,\infty)$). This assumption can be replaced by the weaker assumptions $\sqrt{n}h_0 \underset{n\to\infty}{\longrightarrow} \infty$ and $\sqrt{n}h_\sigma \to \infty$. In this case, we obtain

$$\sup_{x\in D} |\widehat{\sigma}(x) - \sigma(x)| = O_P\left(h_\sigma^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right) + O_P\left(h_0^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right)$$

as a uniform bound for the variance estimator. Assuming that we undersmooth in the two prior fittings, namely $\sqrt{nh}h_\sigma^{p+1} \underset{n\to\infty}{\longrightarrow} 0$ and $\sqrt{nh}h_0^{p+1} \underset{n\to\infty}{\longrightarrow} 0$, but that simultaneously the bandwidths are also large enough to satisfy $\log(n)\sqrt{h}/(\sqrt{n}h_0) \underset{n\to\infty}{\longrightarrow} 0$ and $\log(n)\sqrt{h}/(\sqrt{n}h_\sigma) \underset{n\to\infty}{\longrightarrow} 0$, the two prior fittings will be asymptotically negligible in comparison to the main fitting. Consequently, Theorem 4.12 remains valid under this bandwidth assumptions.

Assumption 3 (i) ensures a Taylor expansion of the variance function of the desired order, whereas Assumptions 3 (ii), 3 (iii), and 3 (v) are similar to previous assumptions.

[1] Note that since $\delta < 2.5$ and $p \geq 1$ it holds $\sqrt{n}h_0 = O\left(n^{\frac{2p+3-2\delta}{4p+6}}\right) \underset{n\to\infty}{\longrightarrow} \infty$. The remaining assumptions can be verified in the same manner as we did in Chapter 3.

4.3 Convergence results for the variance estimator

This section addresses the asymptotic study of the residual based variance estimator. We provide pointwise and uniform rates for this estimation technique.

The problem of estimating the conditional variance function

$$\sigma^2(x) := \mathrm{Var}(Y|X = x)$$

has attracted a lot of research in the last decades, providing manifold estimation approaches. Two main approaches were conducted by Härdle & Tsybakov (1997), who introduced a direct estimator, and Gasser et al. (1986), suggesting difference-based estimation. More recent works concerning the problem of estimating the conditional variance are, for instance, Ziegelmann (2002), who proposed a local linear exponential tilting estimator to ensure its positivity, as well as Linton & Xiao (2007) and Jin et al. (2015) who addressed the problem of an estimator that adapts to the error distribution. Vilar-Fernández & Francisco-Fernández (2006) developed an estimator for the variance function in the fixed-design case and provided asymptotic results.

However, all of this approaches are either not suitable for our purposes in their theoretical performance or require multiple additional assumptions on the error process, the design process or the conditional mean. We therefore consider a residual-based estimator of the variance function that suits our proposed efficient estimation procedure very well. Note that for the squared residuals we have

$$\sigma^2(x) = \mathbb{E}\left((Y - m(X))^2|X = x\right).$$

If the regression function $m(\cdot)$ was known, we could thus apply local polynomial regression on $(Y_t - m(X_t))^2, t = 1, \ldots, n$ to estimate $\sigma^2(x)$. Since $m(\cdot)$ is not known, we substitute its values by estimates obtained via prior smoothing, namely $\widehat{m}(X_t), t = 1, \ldots, n$. Note that the prewhitening procedure requires a prior smoothing anyway. The additional effort of estimating the variance via the residual-based technique is thus comparatively low. Further, as we shall see in Theorems 4.7 and 4.12, the asymptotic behavior of an estimator constructed in this way matches our needs.

The idea of residual-based estimation of the variance function is not novel. Hall & Caroll (1989) considered this approach for kernel[2] estimators and derived pointwise asymptotic rates, which they proved to be asymptotically optimal. Neumann (1994) added a data-driven bandwidth choice to this estimation procedure. A few years later, Ruppert et al. (1997) and Fan & Yao (1998) adapted the residual-based approach to local polynomial smoothing and proposed two-step approaches

[2]Note that in this context, the definition of a kernel function differs from that used in this thesis. Here, a kernel $k(\cdot)$ of order l is defined as a smooth function with compact support, e.g. on $[-1, 1]$ that satisfies $\int k(x)dx = 1$, and $\int x^i k(x)dx = 0$, for all $i = 1, \ldots, l$.

similar to the one used here. Further, they provided pointwise convergence rates for the resulting estimators.

However, the asymptotic analysis of a feasible efficient estimator in the heteroscedastic setting requires us to develop convergence rates that hold uniformly over compact intervals. In what follows, we thus extend the existing results and provide uniform convergence rates for the residual based variance estimator in Theorem 4.7. We obtain the same convergence results as if the regression function $m(\cdot)$ was known: the convergence rate of the bias term B_x^σ is $O_P\left(h_\sigma^{p+1}\right)$ whereas the variance term V_x^σ converges with the usual pointwise rate of $O_P\left(1/\sqrt{nh_\sigma}\right)$, and with a uniform convergence rate of $O_P\left(\sqrt{\log(n)}/\sqrt{nh_\sigma}\right)$.

4.3.1 Pointwise convergence rate

In this section, we derive a pointwise convergence rate for the variance estimator. For this purpose, we first establish a uniform convergence rate for the local polynomial. This rate contributes to the order of the variance estimator via the prior estimation. One technical core for the derivation of uniform convergence results is the following Bernstein-type inequality for strongly mixing processes:

Lemma 4.2 (Bernstein inequality for $\boldsymbol{\alpha}$-mixing processes, Merlevède et al. (2009))**.** *Let $(Z_t)_{t\in\mathbb{Z}}$ be a sequence of stationary, centered α - mixing random variables with exponentially decreasing mixing coefficients. Further, let $\sup_{t\in\mathbb{Z}}\|Z_t\|_\infty \leq M$.*

Then there is a positive constant C_Z depending only on the mixing-rate of $(Z_t)_{t\in\mathbb{Z}}$, such that for all $n \geq 2$, the following inequality for the mean $S_n := \frac{1}{n}\sum_{i=1}^n Z_i$ holds true:

$$\mathrm{P}\left(|S_n| \geq \lambda\right) \leq \exp\left(-\frac{C_Z\lambda^2}{v^2 n + M^2 + \lambda M \log(n)^2}\right),$$

where v^2 is defined via

$$v^2 := \mathrm{Var}(Z_1) + 2\sum_{i=2}^{\infty}|\mathrm{Cov}(Z_1, Z_i)|.$$

Proof. See (Merlevède et al. 2009, Theorem 2). □

With the help of this inequality, we can now derive a pointwise convergence rate for the conventional local polynomial estimator:

Lemma 4.3 (Uniform convergence rate for the conventional local polynomial estimator over compact intervals)**.** *Let D be a compact interval. Then under Assumptions 1 and 3 it holds*

$$\sup_{x\in D}|\widehat{m}(x) - m(x)| = O_P\left(h_0^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right).$$

Proof. Since D is compact, it can be covered by a finite number $L = L(n)$ of intervals $I_k = I_{n,k}$ with length l_n and centers $x_k = x_{n,k}$. Clearly, $l_n = C/L(n)$. We write

$$\begin{aligned}
&\sup_{x\in D} |\widehat{m}(x) - m(x)| \\
&= \max_{1\le k\le L(n)} \sup_{x\in D\cap I_k} |\widehat{m}(x) - m(x)| \\
&\le \max_{1\le k\le L(n)} \sup_{x\in D\cap I_k} |\widehat{m}(x) - \widehat{m}(x_k)| + \max_{1\le k\le L(n)} |\widehat{m}(x_k) - m(x_k)| \\
&\quad + \max_{1\le k\le L(n)} \sup_{x\in D\cap I_k} |m(x_k) - m(x)| \\
&=: Q_{n,1} + Q_{n,2} + Q_{n,3}.
\end{aligned}$$

We show that $Q_{n,1} = O(l_n) + O_P\left(h_0^{p+1} + \frac{l_n n^{1/8}}{h_0^2}\right)$ and $Q_{n,3} = O(l_n)$ as direct consequences from the assumptions. The term $Q_{n,2}$ requires a more profound study. We show that if $L(n) = n^\gamma$ with $\gamma < \infty$, it holds $Q_{n,2} = O_P\left(h_0^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right)$. Set $l_n = \frac{\sqrt{\log(n)}}{\sqrt{nh_0}} \frac{h_0^2}{n^{1/8}}$ (which implies that indeed $L(n) = n^\gamma$), to get

$$\sup_{x\in D} |\widehat{m}(x) - m(x)| = O_P\left(h_0^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right).$$

For $Q_{n,3}$, note that by Assumption 1 (i), $m(\cdot)$ is continuously differentiable with bounded derivatives and hence Lipschitz. Therefore,

$$\max_{1\le k\le L(n)} \sup_{x\in D\cap I_k} |m(x_k) - m(x)| \le C\, l_n.$$

The term $Q_{n,1}$ may be decomposed as follows

$$\begin{aligned}
&\max_{1\le k\le L(n)} \sup_{x\in D\cap I_k} |\widehat{m}(x) - \widehat{m}(x_k)| \\
&\le \max_{1\le k\le L(n)} \sup_{x\in D\cap I_k} \left(|m(x) - m(x_k)| + |B_x - B_{x_k}| + |V_x - V_{x_k}|\right).
\end{aligned}$$

Corollary 2.16 provides $\sup_{x\in D} B_x = O_P(h_0^{p+1})$. For the variance-type term, Corollary 2.15 gives $(\mathbf{X}_0^\mathsf{T}\mathbf{W}_0\mathbf{X}_0)^{-1} \xrightarrow{\mathrm{P}} \mathbf{M}_0^{-1} f_X^{-1}(x)$ uniformly in $x \in D$, where $\mathbf{X}_0$, $\mathbf{W}_0$ and $\mathbf{M}_0$ as $\mathbf{X}$, $\mathbf{W}$ and $\mathbf{M}$ but with kernel $K_0(\cdot)$ and bandwidth h_0. Therefore, we have uniformly in $x \in D$

$$V_x - V_{x_k} = (1 + o_P(1))\, \mathbf{e}_1^\mathsf{T}\mathbf{M}_0{}^{-1} \left(f_X^{-1}(x_k)\mathbf{X}_{0,x_k}^\mathsf{T}\mathbf{W}_{0,x_k} - f_X^{-1}(x)\, \mathbf{X}_0^\mathsf{T}\mathbf{W}_0\right)\mathbf{U},$$

with $\mathbf{X}_{0,x_k}$, $\mathbf{W}_{0,x_k}$ defined as $\mathbf{X}_0$ and $\mathbf{W}_0$, replacing x by x_k. Since $\mathbf{M}_0$ is a regular $(p+1)\times$

$(p+1)-$matrix, it thus suffices to study the order of the following term:

$$\max_{1\le k\le L(n)} \sup_{x\in D\cap I_k} \left|\left(f_X^{-1}(x_k)\left[\mathbf{X}_{0,x_k}^\mathsf{T}\mathbf{W}_{0,x_k}\mathbf{U}\right]_r - f_X^{-1}(x)\ \left[\mathbf{X}_0^\mathsf{T}\mathbf{W}_0\mathbf{U}\right]_r\right)\right|$$
$$\le \max_{1\le k\le L(n)} \sup_{x\in D\cap I_k} \sum_{i=1}^n \left|\left(f_X^{-1}(x_k)\frac{1}{nh_0}K_0\left(\frac{X_i-x_k}{h_0}\right)\left(\frac{X_0-x_k}{h_0}\right)^{r-1}\right.\right.$$
$$\left.\left. -f_X^{-1}(x)\frac{1}{nh_0}K_0\left(\frac{X_i-x}{h_0}\right)\left(\frac{X_i-x}{h_0}\right)^{r-1}\right)U_i\right|,$$

for $r\in\{1,\dots,p+1\}$ arbitrary.

The Lipschitz and boundedness assumptions on the kernel function $K_0(\cdot)$ (Assumption 3 (v)) and the density function $f_X(\cdot)$ (Assumption 3 (ii)) provide

$$\left|f_X^{-1}(x_k)\frac{1}{h_0}K_0\left(\frac{X_i-x_k}{h_0}\right)\left(\frac{X_i-x_k}{h_0}\right)^{r-1} - f_X^{-1}(x)\frac{1}{h_0}K_0\left(\frac{X_i-x}{h_0}\right)\left(\frac{X_i-x}{h_0}\right)^{r-1}\right|$$
$$\le \frac{1}{h_0}C\left|\frac{x-x_k}{h_0}\right|$$
$$\le C\frac{l_n}{h_0^2},$$

uniformly in $x, x_k\in D$.

Further, we get using Bonferroni's and Markov's inequality

$$\mathrm{P}\left(\max_{i=1,\dots,n}|U_i|>n^\delta\right)\le\sum_{i=1}^n \mathrm{P}\left(|U_i|>n^\delta\right)\le\sum_{i=1}^n\frac{\mathbb{E}\left(|U_i|\right)^M}{n^{\delta M}} = \mathbb{E}(|U_1|)^M n^{1-\delta M}.$$

Since $\mathbb{E}(|U_1|)^9<\infty$, set $M=9$, and $\delta=1/8$ to get $\max\limits_{i=1,\dots,n}|U_i| = O_P\left(n^{1/8}\right)$.

Hence

$$\max_{1\le k\le L(n)} \sup_{x\in D\cap I_k} |\widehat{m}(x)-\widehat{m}(x_k)| = O(l_n)+O_P\left(h_0^{p+1}\right)+O_P\left(\frac{l_n n^{1/8}}{h_0^2}\right).$$

The remaining term $Q_{n,2}$ may be decomposed into

$$\max_{1\le k\le L(n)}|\widehat{m}(x_k)-m(x_k)|\le O_P\left(h_0^{p+1}\right)+\max_{1\le k\le L(n)}|V_{x_k}|,$$

by the uniform convergence results of the bias term.

The main task is to show that $\max_{1\le k\le L(n)}|V_{x_k}| = O_P\left(\frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right)$. Similar to previous calculations, the uniform convergence results for $(\mathbf{X}_0^\top \mathbf{W}_0 \mathbf{X}_0)^{-1}$ give

$$\max_{1 \le k \le L(n)} \left| V_{x_k} - \mathbf{e}_1^\top \mathbf{M}_0{}^{-1} \left(f_X^{-1}(x_k) \mathbf{X}_{0,x_k}^\top \mathbf{W}_{0,x_k} \mathbf{U} \right) \right| = o_P(1). \tag{4.1}$$

Define for some $r \in \{1, \dots, p+1\}$

$$S_n(x_k) := \left[\mathbf{X}_{0,x_k}^\top \mathbf{W}_{0,x_k} \mathbf{U} \right]_r = \frac{1}{nh_0} \sum_{i=1}^{n} K_0 \left(\frac{X_i - x_k}{h_0} \right) \left(\frac{X_i - x_k}{h_0} \right)^{r-1} U_i.$$

In what follows, we show that $\max_{1 \le k \le L(n)} |S_n(x_k)| = O_P\left(\frac{\sqrt{\log(n)}}{\sqrt{nh_0}} \right)$, applying the Bernstein-type inequality for strongly mixing processes (see Lemma 4.2) on this sum term. Since $f_X(\cdot)$ is bounded away from zero, this directly implies $\max_{1 \le k \le L(n)} |V_{x_k}| = O_P\left(\frac{\sqrt{\log(n)}}{\sqrt{nh_0}} \right)$ by equation (4.1), and the regularity and finiteness of the $(p+1) \times (p+1)$-dimensional matrix $\mathbf{M}_0$. We proceed step-wise.

Step 1: Definition of asymptotically equivalent truncated sum term. The Bernstein-type inequality requires centered and bounded summands. Since $S_n(x_k)$ is not bounded, we define the truncated and centered sum term:

$$S_n^t(x_k) := \frac{1}{nh_0} \sum_{i=1}^{n} K_0 \left(\frac{X_i - x_k}{h_0} \right) \left(\frac{X_i - x_k}{h_0} \right)^{r-1} \left(U_i 1_{\{|U_i| \le t_n\}} - \mathbb{E}\left(U_i 1_{\{|U_i| \le t_n\}} \right) \right),$$

with $t_n = n^{1/8}$. Similarly, define the truncated variance-type term $V_{x_k}^t$.

We show that the truncation is asymptotically negligible. For this purpose, we decompose the difference between $S_n(x_k)$ and $S_n^t(x_k)$ as follows:

$$\begin{aligned}
&\left| S_n^t(x_k) - S_n(x_k) \right| \\
\le &\left| \frac{1}{nh_0} \sum_{i=1}^{n} K_0 \left(\frac{X_i - x_k}{h_0} \right) \left(\frac{X_i - x_k}{h_0} \right)^{r-1} U_i 1_{\{|U_i| > t_n\}} \right| \\
&+ \left| \frac{1}{nh_0} \sum_{i=1}^{n} K_0 \left(\frac{X_i - x_k}{h_0} \right) \left(\frac{X_i - x_k}{h_0} \right)^{r-1} \mathbb{E}\left(U_i 1_{\{|U_i| > t_n\}} \right) \right| \\
=: &\, I_1(x_k) + I_2(x_k).
\end{aligned}$$

The assertion follows by showing that $\mathrm{P}\left(\max_{1 \le k \le L(n)} I_1(x_k) \neq 0 \right) \underset{n \to \infty}{\longrightarrow} 0$, and $\max_{1 \le k \le L(n)} I_2(x_k) = o\left(\frac{\sqrt{\log(n)}}{\sqrt{nh_0}} \right)$.

Since $(U_t)_{t\in\mathbb{Z}}$ is stationary, we have for $I_1(x_k)$

$$\mathrm{P}\left(\max_{1\le k\le L(n)} I_1(x_k)\neq 0\right)\le n\,\mathrm{P}(|U_1|>t_n).$$

Since the ninth moments of the error process are bounded, Markov's inequality gives

$$\mathrm{P}(|U_1|>t_n)\le \frac{\mathrm{I\!E}\left(|U_1|^9\right)}{t_n^9}=O\left(n^{-9/8}\right),$$

and thus

$$\mathrm{P}\left(\max_{1\le k\le L(n)} I_1(x_k)\neq 0\right)\underset{n\to\infty}{\longrightarrow} 0.$$

For $I_2(x_k)$, Hölder's inequality provides

$$\mathrm{I\!E}\left(|U_i|1_{\{|U_i|>t_n\}}\right)\le (\mathrm{I\!E}|U_i|^9)^{1/9}\left(\frac{\mathrm{I\!E}|U_i|^9}{t_n^9}\right)^{8/9}=O\left(\frac{1}{t_n^8}\right)=O\left(\frac{1}{n}\right).$$

Since the kernel is bounded with compact support, it holds

$$\max_{1\le k\le L(n)}\left|K_0\left(\frac{X_i-x_k}{h_0}\right)\left(\frac{X_i-x_k}{h_0}\right)^{r-1}\right|\le C.$$

By Assumption 3 (vi) on the bandwidths we thus obtain

$$\max_{1\le k\le L(n)} I_2(x_k)=O\left(\frac{1}{nh_0}\right)=o\left(\frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right).$$

Step 2: Application of Bernstein's inequality on truncated sum term. Denote the summands of the truncated sum term

$$Z_{n,i}:=\frac{1}{nh_0}K_0\left(\frac{X_i-x_k}{h_0}\right)\left(\frac{X_i-x_k}{h_0}\right)^{r-1}\left(U_i1_{\{|U_i|\le t_n\}}-\mathrm{I\!E}\left(U_i1_{\{|U_i|\le t_n\}}\right)\right).$$

The truncation and the boundedness of the kernel $K_0(\cdot)$ give $|Z_{n,i}|\le 2M_{K_0}\frac{1}{nh_0}t_n$, for all $i=1,\dots,n$, with M_{K_0} a constant bounding $K_0(\cdot)$.

Since the processes $(X_t)_{t\in\mathbb{Z}}$ and $(U_t)_{t\in\mathbb{Z}}$ are strongly mixing with exponentially decreasing mixing coefficients, so are the two factors in $(Z_{n,i})_{i\ge1}$ with the same mixing coefficients $\alpha_X(\cdot)$ and $\alpha_U(\cdot)$ respectively. Since $(X_t)_{t\in\mathbb{Z}}$ and $(U_t)_{t\in\mathbb{Z}}$ are further independent, it follows from Lemma 2.7 that the stationary sequence $(Z_{n,i})_{i\ge1}$ is also α-mixing with exponentially decreasing mixing coefficients $\alpha_Z(k)\le\alpha_X(k)+\alpha_U(k)$.

We can thus apply the Bernstein-type inequality (Lemma 4.2) on $S_n^t(x_k)$ with $\lambda:=\eta\frac{\sqrt{\log(n)}}{\sqrt{nh_0}}$ for some arbitrary $\eta > 0$ and $M := M_{K_0} \frac{1}{nh_0} t_n = \sup_i \|Z_{n,i}\|_\infty$ to get

$$\mathrm{P}\left(|S_n^t(x_k)| \geq \eta \, \frac{\sqrt{\log(n)}}{\sqrt{nh_0}} \right) \leq \exp\left(-\frac{C_Z \, \eta^2 \, \frac{\log(n)}{nh_0}}{v^2 n + M^2 + \eta \frac{\sqrt{\log(n)}}{\sqrt{nh_0}} M \log(n)^2} \right),$$

with C_Z a positive constant only depending on the decay rate of $\alpha_Z(\cdot)$, and

$$v^2 := \mathrm{Var}\left(Z_{n,1}\right) + 2 \sum_{i=2}^{\infty} |\mathrm{Cov}(Z_{n,1}, Z_{n,i})|.$$

In the next step we show that $v^2 = O\left(n^{-2} h_0^{-1}\right)$. Further, note that

$$\begin{aligned} M^2 + \eta \, \frac{\sqrt{\log(n)}}{\sqrt{nh_0}} M \log(n)^2 &= M_{K_0}^2 \frac{1}{n^2 h_0^2} t_n^2 + \eta \, \frac{\sqrt{\log(n)}}{\sqrt{nh_0}} M_{K_0} \frac{1}{nh_0} t_n (\log(n))^2 \\ &= M_{K_0}^2 \frac{n^{1/4}}{(nh_0)^2} + \eta \, (\log(n))^{5/2} M_{K_0} \frac{n^{1/8}}{(nh_0)^{3/2}} \\ &= o\left(\frac{1}{nh_0} \right), \end{aligned}$$

since $n^{1/2} h_0 \underset{n \to \infty}{\longrightarrow} \infty$ by Assumption 3 (vi) on the bandwidths.

We therefore obtain for n sufficiently large

$$\begin{aligned} \mathrm{P}\left(|S_n^t(x_k)| \geq \eta \frac{\sqrt{\log(n)}}{\sqrt{nh_0}} \right) &\leq \exp\left(-\frac{C_Z \, \eta^2 \, \frac{\log(n)}{nh_0}}{O\left(\frac{1}{nh_0}\right) + o\left(\frac{1}{nh_0}\right)} \right) \\ &\leq \exp\big(- C\eta^2 \log(n) \big) \\ &= n^{-\eta^2 C}. \end{aligned}$$

Since $L(n) = n^\gamma$, we thus get from Bonferroni's inequality

$$\mathrm{P}\left(\max_{1 \leq k \leq L(n)} |S_n^t(x_k)| \geq \eta \frac{\sqrt{\log(n)}}{\sqrt{nh_0}} \right) \leq L(n) n^{-\eta^2 C} = n^{\gamma - \eta^2 C} \underset{n \to \infty}{\longrightarrow} 0,$$

by choosing η large enough.

Hence $\max_{1 \leq k \leq L(n)} |S_n^t(x_k)| = O_P\left(\frac{\sqrt{\log(n)}}{\sqrt{nh_0}} \right)$ and since the truncation is asymptotically negligible we obtain

$$\max_{1 \leq k \leq L(n)} |S_n(x_k)| = O_P\left(\frac{\sqrt{\log(n)}}{\sqrt{nh_0}} \right).$$

Step 3: Asymptotic order of the term v^2. It remains to show that

$$v^2 = \mathrm{Var}\left(Z_{n,1}\right) + 2\sum_{i=2}^{\infty} |\mathrm{Cov}(Z_{n,1}, Z_{n,i})| = O\left(\frac{1}{n^2 h_0}\right).$$

Since $(X_t)_{t\in\mathbb{Z}}$ and $(U_t)_{t\in\mathbb{Z}}$ are independent, it holds

$$\begin{aligned}
&\mathrm{Var}\left(Z_{n,1}\right) \\
&= \mathbb{E}\left(\frac{1}{nh_0}K_0\left(\frac{X_1 - x_k}{h_0}\right)\left(\frac{X_1 - x_k}{h_0}\right)^{r-1}\left(U_1 1_{\{|U_1|\le t_n\}} - \mathbb{E}\left(U_1 1_{\{|U_1|\le t_n\}}\right)\right)\right)^2 \\
&\le C\,\mathbb{E}\left(\frac{1}{nh_0}K_0\left(\frac{X_1 - x_k}{h_0}\right)\left(\frac{X_1 - x_k}{h_0}\right)^{r-1}\right)^2 \\
&= O\left(\frac{1}{n^2 h_0}\right),
\end{aligned} \tag{4.2}$$

where we made use of the fact, that the kernel is bounded with compact support, and

$$\begin{aligned}
\mathbb{E}\left(\left(U_1 1_{\{|U_1|\le t_n\}} - \mathbb{E}\left(U_1 1_{\{|U_1|\le t_n\}}\right)\right)^2\right) &\le \mathbb{E}\left(U_1^2 1_{\{|U_1|\le t_n\}}\right) + \left(\mathbb{E}\left(U_1 1_{\{|U_1|\le t_n\}}\right)\right)^2 \\
&\le \mathbb{E}\left(U_1^2\right) + \left(\mathbb{E}\left(U_1\right)\right)^2 \\
&\le C,
\end{aligned}$$

by Assumption 3 (iii) on the error moments.

For the sum of covariance terms, we proceed in complete analogy to the proof of Lemma 2.9. We decompose

$$\sum_{i=2}^{\infty} |\mathrm{Cov}(Z_{n,1}, Z_{n,i})| = \sum_{2\le i\le h_0^{-7/9}} |\mathrm{Cov}(Z_{n,1}, Z_{n,i})| + \sum_{i\ge h_0^{-7/9}} |\mathrm{Cov}(Z_{n,1}, Z_{n,i})|.$$

Similar to the calculations for $\mathrm{Var}(Z_{n,1})$, we obtain $|\mathrm{Cov}(Z_{n,1}, Z_{n,i})| = O\left(\frac{1}{n^2}\right)$. Thus it holds for the first term

$$\sum_{2\le i\le h_0^{-7/9}} |\mathrm{Cov}(Z_{n,1}, Z_{n,i})| = O\left(\frac{h_0^{-7/9} - 1}{n^2}\right) = o\left(\frac{1}{n^2 h_0}\right). \tag{4.3}$$

We now show that the second sum term is of order $O\left(\frac{1}{n^2 h_0}\right)$, which yields the rate of v^2. Set $\widetilde{Z}_{n,i} := nZ_{n,i}$, and note that $|\mathrm{Cov}(\widetilde{Z}_{n,1}, \widetilde{Z}_{n,i})| = n^2|\mathrm{Cov}(Z_{n,1}, Z_{n,i})|$. Since $\left(\widetilde{Z}_{n,i}\right)_{i\ge 1}$ is strongly mixing as shown in the first step, the covariance inequality for strongly mixing processes

(Lemma 2.5) provides

$$\left|\operatorname{Cov}\left(\widetilde{Z}_{n,1}, \widetilde{Z}_{n,i}\right)\right| \leq 4(\alpha_Z(i-1))^{7/9} \left\|\widetilde{Z}_{n,1}\right\|_9^2.$$

Since design and error process are independent, we obtain

$$\begin{aligned}
&\mathbb{E}\left(\left|\widetilde{Z}_{n,1}\right|^9\right) \\
&= \mathbb{E}\left(\left|U_1 1_{\{|U_1| \leq t_n\}} - \mathbb{E}\left(U_1 1_{\{|U_1| \leq t_n\}}\right)\right|^9\right) \mathbb{E}\left(\left|\frac{1}{h_0} K_0\left(\frac{X_1 - x_k}{h_0}\right)\left(\frac{X_1 - x_k}{h_0}\right)^{r-1}\right|^9\right).
\end{aligned}$$

Since the kernel $K_0(\cdot)$ is finite with compact support, it holds

$$\mathbb{E}\left(\left|\frac{1}{h_0} K_0\left(\frac{X_1 - x_k}{h_0}\right)\left(\frac{X_1 - x_k}{h_0}\right)^{r-1}\right|^9\right) \leq \frac{C}{h_0^8}.$$

For the error process, the binomial theorem provides

$$\begin{aligned}
&\mathbb{E}\left(\left|U_1 1_{\{|U_1| \leq t_n\}} - \mathbb{E}\left(U_1 1_{\{|U_1| \leq t_n\}}\right)\right|^9\right) \\
&\leq \mathbb{E}\left(\left(\left|U_1 1_{\{|U_1| \leq t_n\}}\right| + \left|\mathbb{E}\left(U_1 1_{\{|U_1| \leq t_n\}}\right)\right|\right)^9\right) \\
&\leq \sum_{k=0}^{9} \binom{9}{k} \mathbb{E}\left(|U_1|^{9-k} 1_{\{|U_1| \leq t_n\}}\right) \left|\mathbb{E}\left(U_1 1_{\{|U_1| \leq t_n\}}\right)\right|^k \\
&\leq C,
\end{aligned}$$

since the ninth moment of U_1 is finite by Assumption 3 (iii). Hence $\mathbb{E}\left(\left|\widetilde{Z}_{n,1}\right|^9\right) \leq C h_0^{-8}$, and thus $\left|\operatorname{Cov}\left(\widetilde{Z}_{n,1}, \widetilde{Z}_{n,i}\right)\right| \leq C\left(\alpha_Z(i-1)\right)^{7/9} h_0^{-16/9}$.

Consequently, we have for the second sum term

$$\begin{aligned}
\sum_{i \geq h_0^{-7/9}} |\operatorname{Cov}(Z_{n,1}, Z_{n,i})| &\leq \frac{C}{n^2 h_0^{16/9}} \sum_{i \geq h_0^{-7/9}} (\alpha_Z(i-1))^{7/9} \\
&\leq \frac{C}{n^2 h_0^{16/9} h_0^{-7/9}} \sum_{i \geq h_0^{-7/9}} i\left(\alpha_Z(i-1)\right)^{7/9} \qquad (4.4) \\
&= O\left(\frac{1}{n^2 h_0}\right),
\end{aligned}$$

since the mixing coefficients $\alpha_Z(\cdot)$ are exponentially decreasing. Equations (4.2)–(4.4) yield the asymptotic order of v^2. □

We can now provide a pointwise convergence rate for the variance estimator:

Lemma 4.4 (Pointwise convergence rate for the variance estimator)**.** *Under virtue of Assumptions 1 and 3, it holds*

$$\widehat{\sigma^2}(x) - \sigma^2(x) = O_P(h_\sigma^{p+1}) + O_P\left(\frac{1}{\sqrt{nh_\sigma}}\right).$$

Proof. Since we smooth $\left(\widehat{\sigma(X_t)U_t}\right)^2$ on X_t where the true values are $\sigma^2(X_t)$ for $t = 1, \dots, n$, the underlying regression model can be formulated as follows:

$$\left(\widehat{\sigma(X_t)U_t}\right)^2 = \sigma^2(X_t) + V_t, \qquad t = 1, \dots, n,$$

with the error terms $V_t := \left(\widehat{\sigma(X_t)U_t}\right)^2 - \sigma^2(X_t)$.

As for the estimation of $m(\cdot)$ in Chapter 2, expanding $\sigma^2(X_i)$ in a Taylor series around x provides

$$\sigma^2(X_i) = \sigma^2(x) + \sum_{j=1}^{p} \frac{(\sigma^2)^{(j)}(x)}{j!} h_\sigma^j \left(\frac{X_i - x}{h_\sigma}\right)^j + \frac{(\sigma^2)^{(p+1)}(\xi_i)}{(p+1)!} h_\sigma^{p+1} \left(\frac{X_i - x}{h_\sigma}\right)^{p+1},$$

for some real value ξ_i between x and X_i.

Define $\mathbf{X}_\sigma$, $\mathbf{W}_\sigma$ and $\mathbf{M}_\sigma$ as $\mathbf{X}$, $\mathbf{W}$ and $\mathbf{M}$, but with with bandwidth h_σ and kernel $K_\sigma(\cdot)$, and further $\mathbf{\Sigma}^2 := (\sigma^2(X_1), \dots, \sigma^2(X_n))^\intercal$. We have

$$\mathbf{\Sigma}^2 = \mathbf{X}_\sigma \begin{pmatrix} \sigma^2(x) \\ \vdots \\ \frac{(\sigma^2)^{(p)}(x)}{p!} h_\sigma^p \end{pmatrix} + h_\sigma^{p+1} \underbrace{\begin{pmatrix} \frac{(\sigma^2)^{(p+1)}(\xi_1)}{(p+1)!} \left(\frac{X_1 - x}{h_\sigma}\right)^{p+1} \\ \vdots \\ \frac{(\sigma^2)^{(p+1)}(\xi_n)}{(p+1)!} \left(\frac{X_n - x}{h_\sigma}\right)^{p+1} \end{pmatrix}}_{=:\mathbf{B}_\sigma},$$

Define now the vector of squared residuals $\widehat{\mathbf{R}}^2 := (\widehat{r}_1^2, \dots, \widehat{r}_n^2)^\intercal$, with $\widehat{r}_i^2 := \left(\widehat{\sigma(X_i)U_i}\right)^2 = \left(Y_i - \widehat{m(X_i)}\right)^2$, and the vector of error terms from the residual smoothing $\mathbf{V} := (V_1, \dots, V_n)^\intercal$. Obviously, we have $\widehat{\mathbf{R}}^2 = \mathbf{\Sigma}^2 + \mathbf{V}$, and hence

$$\begin{aligned} \widehat{\sigma^2}(x) &= \mathbf{e}_1^\intercal (\mathbf{X}_\sigma^\intercal \mathbf{W}_\sigma \mathbf{X}_\sigma)^{-1} \mathbf{X}_\sigma^\intercal \mathbf{W}_\sigma \widehat{\mathbf{R}}^2 \\ &= \mathbf{e}_1^\intercal (\mathbf{X}_\sigma^\intercal \mathbf{W}_\sigma \mathbf{X}_\sigma)^{-1} \mathbf{X}_\sigma^\intercal \mathbf{W}_\sigma (\mathbf{\Sigma}^2 + \mathbf{V}) \\ &= \sigma^2(x) + \mathbf{e}_1^\intercal h_\sigma^{p+1} (\mathbf{X}_\sigma^\intercal \mathbf{W}_\sigma \mathbf{X}_\sigma)^{-1} \mathbf{X}_\sigma^\intercal \mathbf{W}_\sigma \mathbf{B}_\sigma + \mathbf{e}_1^\intercal (\mathbf{X}_\sigma^\intercal \mathbf{W}_\sigma \mathbf{X}_\sigma)^{-1} \mathbf{X}_\sigma^\intercal \mathbf{W}_\sigma \mathbf{V} \\ &=: \sigma^2(x) + B_x^\sigma + V_x^\sigma. \end{aligned}$$

The uniform convergence results in Chapter 2 provide $(\mathbf{X}_\sigma^\intercal \mathbf{W}_\sigma \mathbf{X}_\sigma)^{-1} \to f_X^{-1}(x) \mathbf{M}_\sigma^{-1}$ almost surely and similarly, $\mathbf{X}_\sigma^\intercal \mathbf{W}_\sigma \mathbf{B}_\sigma = O_P(1)$ both uniformly in $x \in D$. Since $(\sigma^2)^{(p+1)}(\cdot)$ is bounded by Assumption 3 (i), we have[3]

$$\sup_{x\in D} B_x^\sigma = O_P(h_\sigma^{p+1}).$$

Hence

$$\widehat{\sigma^2}(x) = \sigma^2(x) + O_P(h_\sigma^{p+1}) + \mathbf{e}_1^\top f_X^{-1}(x)\mathbf{M}_\sigma^{-1}\, \mathbf{X}_\sigma^\top \mathbf{W}_\sigma \mathbf{V}(1+o_P(1)).$$

In analogy to the calculations for V_x in previous chapters, since $f_X(\cdot)$ is bounded away from zero, and $\mathbf{M}_\sigma$ is a regular matrix, the order of the variance-type term is verified by studying the asymptotic order of the term $\mathbf{X}_\sigma^\top \mathbf{W}_\sigma \mathbf{V}$. In what follows, we show that this term converges with an asymptotic rate of $O_P\left(h_\sigma^{p+1} + \frac{1}{\sqrt{nh_\sigma}}\right)$.[4] We decompose the difference between the squared residuals and the variance function as follows:

$$\begin{aligned}
\widehat{r}_i^2 - \sigma^2(X_i) &= (Y_i - \widehat{m}(X_i))^2 - \sigma^2(X_i) \\
&= (Y_i - m(X_i) + m(X_i) - \widehat{m}(X_i))^2 - \sigma^2(X_i) \\
&= \sigma^2(X_i)(U_i^2 - 1) + 2\sigma(X_i)U_i(m(X_i) - \widehat{m}(X_i)) + (m(X_i) - \widehat{m}(X_i))^2.
\end{aligned}$$

We thus have

$$\begin{aligned}
\left[\mathbf{X}_\sigma^\top \mathbf{W}_\sigma \mathbf{V}\right]_r =& \sum_{i=1}^n \frac{1}{nh_\sigma} K_\sigma\left(\frac{X_i - x}{h_\sigma}\right)\left(\frac{X_i - x}{h_\sigma}\right)^{r-1} \left(\sigma^2(X_i)(U_i^2 - 1)\right. \\
&+ \left. 2\sigma(X_i)U_i(m(X_i) - \widehat{m}(X_i)) + (m(X_i) - \widehat{m}(X_i))^2\right).
\end{aligned}$$

In what follows, we show that

$$\begin{aligned}
S_{n,1} :=& \frac{1}{nh_\sigma}\sum_{i=1}^n C_{K_\sigma}(X_i)\, \sigma^2(X_i)(U_i^2 - 1) = O_P\left(\frac{1}{\sqrt{nh_\sigma}}\right), \\
S_{n,2} :=& \frac{1}{nh_\sigma}\sum_{i=1}^n C_{K_\sigma}(X_i)\, 2\sigma(X_i)U_i(m(X_i) - \widehat{m}(X_i)) = O_P\left(h_0^{p+1}\right) + O_P\left(\frac{1}{nh_0}\right), \\
S_{n,3} :=& \frac{1}{nh_\sigma}\sum_{i=1}^n C_{K_\sigma}(X_i)\, (m(X_i) - \widehat{m}(X_i))^2 = O_P\left(h_0^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right)^2,
\end{aligned}$$

with $C_{K_\sigma}(y) := K_\sigma\left(\frac{y-x}{h_\sigma}\right)\left(\frac{y-x}{h_\sigma}\right)^{r-1}$. Thus by Assumption 3 (vi) on the bandwidths

[3] At this stage we do not need a uniform rate for the bias. However, we will make use of it later on when deriving uniform rates for the variance estimator.

[4] Note that we have a bias term due to the previous local polynomial smoothing.

$\left[\mathbf{X}_\sigma^\mathsf{T}\mathbf{W}_\sigma\mathbf{V}\right]_r = O_P\left(h_\sigma^{p+1} + \frac{1}{\sqrt{nh_\sigma}}\right)$ and consequently

$$V_x^\sigma = O_P\left(h_\sigma^{p+1} + \frac{1}{\sqrt{nh_\sigma}}\right).$$

It remains to calculate the convergence rate of the $S_{n,i}$ for $i = 1,2,3$. We start with $S_{n,1}$. We get by the stationarity and independence assumptions on $(X_t)_{t\in\mathbb{Z}}$ and $(U_t)_{t\in\mathbb{Z}}$ and since $\mathbb{E}\left(C_{K_\sigma}(X_i)(U_i^2-1)\right) = 0$, that

$$\begin{aligned}
\operatorname{Var}(S_{n,1}) \leq& \frac{1}{nh_\sigma^2}\operatorname{Var}\left(C_{K_\sigma}(X_1)\sigma^2(X_1))(U_1^2-1)\right) \\
&+ \frac{2}{nh_\sigma^2}\sum_{k=2}^{n}\left|\operatorname{Cov}\left(C_{K_\sigma}(X_1)\sigma^2(X_1)\left(U_1^2-1\right), C_{K_\sigma}(X_k)\sigma^2(X_k)\left(U_k^2-1\right)\right)\right| \\
=& \frac{1}{nh_\sigma^2}\mathbb{E}\left(C_{K_\sigma}^2(X_1)\sigma^2(X_1))\right)\mathbb{E}\left((U_1^2-1)^2\right) \\
&+ \frac{2}{nh_\sigma^2}\sum_{k=2}^{n}\mathbb{E}\left(C_{K_\sigma}(X_1)\sigma^2(X_1)C_{K_\sigma}(X_k)\sigma^2(X_k)\right)\mathbb{E}\left|\left(U_1^2-1\right)\left(U_k^2-1\right)\right| \\
=& O\left(\frac{1}{nh_\sigma}\right) + O\left(\frac{1}{n}\right),
\end{aligned}$$

since by Assumption 3 (v) and 3 (ii) on the kernel function and the variance function, we have $\mathbb{E}\left(C_{K_\sigma}^2(X_1)\sigma^2(X_1))\right) = O\left(h_\sigma\right)$, $\mathbb{E}\left(C_{K_\sigma}(X_1)\sigma^2(X_1)C_{K_\sigma}(X_k)\sigma^2(X_k)\right) = O(h_\sigma^2)$. Further, Assumption 3 (iii) on the error process provides $\mathbb{E}\left((U_1^2-1)^2\right) \leq C$, and

$$\sum_{k=2}^{n}\left|\mathbb{E}\left(\left(U_1^2-1\right)\left(U_k^2-1\right)\right)\right| = O(1),$$

since Lemma 3.6 for the summed covariance holds true for $(1-U_t^2)_{t\in\mathbb{Z}}$.[5] Hence, $S_{n,1} = O_P\left(\frac{1}{\sqrt{nh_\sigma}}\right)$ by Chebyshev's inequality.

The uniform convergence results for the bias term $B_x = O_P\left(h_0^{p+1}\right)$ provide

$$S_{n,2} = O_P\left(h_0^{p+1}\right) + \sum_{i=1}^{n}\frac{1}{nh_\sigma}C_{K_\sigma}(X_i)2\sigma(X_i)U_iV_{X_i}.$$

As shown in Chapter 3, it holds $\left|V_x - \mathbf{e}_1^\mathsf{T}\mathbf{M}_0^{-1}f_X(x)^{-1}\mathbf{X}_0^T\mathbf{W}_0\mathbf{U}\right| = o_P(1)$ uniformly in $x \in D$, with $\mathbf{M}_0^{-1}$ a finite $(p+1)\times(p+1)$-matrix. Keeping in mind that $\sigma(\cdot)$ is bounded and $f_X(\cdot)$ is bounded away from zero, it thus suffices to study the order of the following term:

$$\widetilde{S}_{n,2} = \sum_{i=1}^{n}\frac{1}{nh_\sigma}C_{K_\sigma}(X_i)\frac{1}{nh_0}\sum_{j=1}^{n}C_{K_0}(X_j,X_i)U_iU_j,$$

[5] As a function of $(U_t)_{t\in\mathbb{Z}}$, $(1-U_t^2)_{t\in\mathbb{Z}}$ is $\alpha-$mixing with the same exponentially decreasing mixing coefficients.

with $C_{K_0}(y,z) := K_0\left(\frac{y-z}{h_0}\right)\left(\frac{y-z}{h_0}\right)^{s-1}$, for some $s \in \{1,\dots,p+1\}$ arbitrary. It holds

$$\begin{aligned}
&\mathbb{E}(\widetilde{S}^2_{n,2}) \\
&= \frac{1}{n^4 h_\sigma^2 h_0^2} \sum_{i,j,l,m} \mathbb{E}(C_{K_\sigma}(X_i)C_{K_0}(X_j,X_i)C_{K_\sigma}(X_l)C_{K_0}(X_m,X_l))\mathbb{E}(U_iU_jU_lU_m).
\end{aligned}$$

As in the proof Lemma 3.9 we have $\sum_{i,j,l,m} \mathbb{E}(U_iU_jU_lU_m) = O(n^2)$. Further, since the kernel functions are bounded with compact support by Assumption 3 (iii), we obtain

$$\begin{aligned}
&\mathbb{E}(C_{K_\sigma}(X_i)C_{K_0}(X_j,X_i)C_{K_\sigma}(X_l)C_{K_0}(X_m,X_l)) \\
&\leq C\,\mathbb{E}(C_{K_\sigma}(X_i)C_{K_0}(X_j,X_i)) = O(h_0 h_\sigma).
\end{aligned}$$

Hence $\mathbb{E}\left(\widetilde{S}^2_{n,2}\right) = O\left(\frac{1}{n^2 h_\sigma h_0}\right)$, and thus by Markov's inequality $\widetilde{S}_{n,2} = O_P\left(\frac{1}{nh_0}\right)$.

For $S_{n,3}$, we have

$$|S_{n,3}| \leq \sup_x (\widehat{m}(x) - m(x))^2 \frac{1}{nh_\sigma} \sum_{i=1}^n |C_{K_\sigma}(X_l)|$$

The uniform convergence result for the local polynomial estimator (Lemma 4.3) provides $\sup_x(\widehat{m}(x) - m(x)) = O_P\left(h_0^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right)$. Markov's inequality applied on the remaining term concludes the proof. ☐

Corollary 4.5. *Under Assumptions 1 and 3 we have*

$$\sigma(x) - \widehat{\sigma}(x) = O_P(h_\sigma^{p+1}) + O_P\left(\frac{1}{\sqrt{nh_\sigma}}\right).$$

Proof. Since $\widehat{\sigma}(x) > 0$ by definition, it holds

$$|\sigma(x) - \widehat{\sigma}(x)| = \frac{\left|\sigma^2(x) - \widehat{\sigma^2}(x)\right|}{\sigma(x) + \widehat{\sigma}(x)} \leq \frac{1}{C_{\sigma,1}} \left|\sigma^2(x) - \widehat{\sigma^2}(x)\right|.$$

☐

Remark 4.6. *The proof of Lemma 4.4 provides that the order of the prior smoothing is asymptotically negligible if we undersmooth in this first smoothing. The variance estimator converges with the same rate as if the residuals were known.*

4.3.2 Uniform convergence rate over compact intervals

In this section, we derive a convergence rate for the variance estimator that holds uniformly over compact intervals:

Theorem 4.7 (Uniform convergence rate for the variance estimator over compact intervals)**.** *Let D be a compact interval. Then under Assumptions 1 and 3, it holds*

$$\sup_{x \in D} |\widehat{\sigma^2}(x) - \sigma^2(x)| = O_P\left(h_\sigma^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}} \right).$$

Proof. We take over the basic steps from the proof of Lemma 4.3.

Again we cover D by a finite number $L = L(n)$ of intervals $I_k = I_{n,k}$ with length l_n and centers $x_k = x_{n,k}$, and write

$$\begin{aligned}
&\sup_{x \in D} \left| \widehat{\sigma^2}(x) - \sigma^2(x) \right| \\
&= \max_{1 \le k \le L(n)} \sup_{x \in D \cap I_k} \left| \widehat{\sigma^2}(x) - \sigma^2(x) \right| \\
&\le \max_{1 \le k \le L(n)} \sup_{x \in D \cap I_k} \left| \widehat{\sigma^2}(x) - \widehat{\sigma^2}(x_k) \right| + \max_{1 \le k \le L(n)} \left| \widehat{\sigma^2}(x_k) - \sigma^2(x_k) \right| \\
&\quad + \max_{1 \le k \le L(n)} \sup_{x \in D \cap I_k} \left| \sigma^2(x_k) - \sigma^2(x) \right| \\
&=: Q_{n,1}^\sigma + Q_{n,2}^\sigma + Q_{n,3}^\sigma.
\end{aligned}$$

We show that $Q_{n,1}^\sigma = O(l_n) + O_P\left(h_\sigma^{p+1}\right) + O_P\left(\frac{n^{1/4} l_n}{h_\sigma^2}\right)$ and $Q_{n,3}^\sigma = O(l_n)$ adapting the proof of Lemma 4.3. As for the conventional local polynomial estimator, we show that $Q_{n,2}^\sigma = O_P\left(h_\sigma^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right)$, if $L(n) = n^\gamma$ for some $\gamma < \infty$.

Set $l_n = \frac{\sqrt{\log(n)} h_\sigma^2}{\sqrt{nh_\sigma} n^{1/4}}$ to get

$$\sup_{x \in D} \left| \widehat{\sigma^2}(x) - \sigma^2(x) \right| = O_P\left(h_\sigma^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}} \right).$$

For $Q_{n,3}^\sigma$, we have

$$\sup_{x \in D \cap I_k} \left| \sigma^2(x_k) - \sigma^2(x) \right| \le C l_n,$$

since $\sigma^2(\cdot)$ is continuously differentiable and thus local Lipschitz continuous.

Decomposition of $Q_{n,1}^\sigma$ provides

$$\begin{aligned}
&\max_{1 \le k \le L(n)} \sup_{x \in D \cap I_k} \left| \widehat{\sigma^2}(x) - \widehat{\sigma^2}(x_k) \right| \\
&= \max_{1 \le k \le L(n)} \sup_{x \in D \cap I_k} \left| (\sigma^2(x) - \sigma^2(x_k)) + (B_x^\sigma - B_{x_k}^\sigma) + (V_x^\sigma - V_{x_k}^\sigma) \right|.
\end{aligned}$$

As in the proof of Lemma 4.4, we have $\sup_{x \in D} B_x^\sigma = O_P\left(h_\sigma^{p+1}\right)$ for the bias term.

For the variance-type term $\left|V_x^\sigma - V_{x_k}^\sigma\right|$, by the uniform convergence results of $(\mathbf{X}_\sigma^\mathsf{T}\mathbf{W}_\sigma\mathbf{X}_\sigma)^{-1}$ it suffices to study the order of

$$\max_{1\le k\le L(n)} \sup_{x\in D\cap I_k} \sum_{i=1}^n \Bigg| \Bigg(f_X^{-1}(x)\frac{1}{nh_\sigma}K_\sigma\left(\frac{X_i-x}{h_\sigma}\right)\left(\frac{X_i-x}{h_\sigma}\right)^{r-1} \\ -f_X^{-1}(x_k)\frac{1}{nh_\sigma}K_\sigma\left(\frac{X_i-x_k}{h_\sigma}\right)\left(\frac{X_i-x_k}{h_\sigma}\right)^{r-1}\Bigg) V_i \Bigg|,$$

with V_i as defined in Lemma 4.4, and $r\in\{1,\dots,p+1\}$ arbitrary.

We show that $\max_{i=1,\dots,n} V_i = O_P\left(n^{1/4}\right)$. A combination of Bonferroni's and Markov's inequality gives

$$\mathrm{P}\left(\max_{i=1,\dots,n} U_i^2 > n^\delta\right) \le \sum_{i=1}^n \frac{\mathbb{E}\left(U_i^{2M}\right)}{n^{\delta M}} \le \mathbb{E}(U_1^{2M})n^{1-\delta M}.$$

Set $M = 9/2$, and $\delta = 1/4$ to get $\max_{i=1,\dots,n} U_i^2 = O_P\left(n^{1/4}\right)$, and consequently $\max_{i=1,\dots,n} |U_i| = O_P\left(n^{1/8}\right)$. Further, the uniform convergence results for the local polynomial estimator give $\sup_{x\in D} |m(x)-\widehat{m}(x)| = O_P\left(h_0^{p+1}\right) + O_P\left(\frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right)$.

Therefore and since $\sigma(\cdot)$ is uniformly bounded,

$$\begin{aligned}
&\max_{i=1,\dots,n} V_i \\
&= \max_{i=1,\dots,n} \left(\sigma^2(X_i)(U_i^2-1) + 2\sigma(X_i)U_i(m(X_i)-\widehat{m}(X_i)) + (m(X_i)-\widehat{m}(X_i))^2\right) \\
&= O_P\left(n^{1/4}\right),
\end{aligned}$$

which provides $\sup_{x\in D\cap I_k} |V_x^\sigma - V_{x_k}^\sigma| = O_P\left(\frac{n^{1/4}l_n}{h_\sigma^2}\right)$ as in the proof of Lemma 4.3. This yields the desired order of $Q_{n,1}^\sigma$.

For the remaining term $Q_{n,2}^\sigma$, the uniform convergence results for the bias term (see the proof of Lemma 4.4) provide

$$\max_{1\le k\le L(n)} \left|\widehat{\sigma^2}(x_k) - \sigma^2(x_k)\right| \le O_P\left(h_\sigma^{p+1}\right) + \max_{1\le k\le L(n)} |V_{x_k}^\sigma|. \tag{4.5}$$

For the remaining variance-type term V_x^σ, again by the uniform convergence results of $(\mathbf{X}_\sigma^\mathsf{T}\mathbf{W}_\sigma\mathbf{X}_\sigma)^{-1}$, and since $f_X(\cdot)$ is bounded away from zero, it suffices to study the order of

$$\max_{1\le k\le L(n)} \left|\left[\mathbf{X}_\sigma^\mathsf{T}\mathbf{W}_\sigma\mathbf{V}\right]_r\right| =: \max_{1\le k\le L(n)} \left|S_{n,1}(x_k) + S_{n,2}(x_k) + S_{n,3}(x_k)\right|, \tag{4.6}$$

with

$$S_{n,1}(x_k) = \frac{1}{nh_\sigma} \sum_{i=1}^{n} C_{K_\sigma}(X_i)\sigma^2(X_i)(U_i^2 - 1),$$
$$S_{n,2}(x_k) = \frac{1}{nh_\sigma} \sum_{i=1}^{n} C_{K_\sigma}(X_i) 2\,\sigma(X_i)U_i(m(X_i) - \widehat{m}(X_i)),$$
$$S_{n,3}(x_k) = \frac{1}{nh_\sigma} \sum_{i=1}^{n} C_{K_\sigma}(X_i)(m(X_i) - \widehat{m}(X_i))^2,$$

where $C_{K_\sigma}(y) := K_\sigma\left(\frac{y-x_k}{h_\sigma}\right)\left(\frac{y-x_k}{h_\sigma}\right)^{r-1}$ for some $r \in \{1, \ldots, p+1\}$ arbitrary.

In Lemmas 4.8 - 4.10 we show that

$$\max_{1\le k\le L(n)} |S_{n,i}(x_k)| = O_P\left(h_\sigma^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right),\ i = 1, 2, 3. \tag{4.7}$$

In detail, we show that $\max_{1\le k\le L(n)} |S_{n,1}(x_k)| = O_P\left(\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right)$ in analogy to the proof of Lemma 4.3. Further, we derive the orders $\max_{1\le k\le L(n)} |S_{n,2}(x_k)| = O_P\left(h_0^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right)$ and $\max_{1\le k\le L(n)} |S_{n,3}(x_k)| = O_P\left(h_0^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right)^2$, where we make use of the uniform convergence results for the local polynomial estimator. Assumption 3 (vi) on the bandwidths yields (4.7).

Equations (4.5) – (4.7) provide $Q_{n,2}^\sigma = O_P\left(h_\sigma^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right)$, which concludes the proof. □

It remains to prove the asymptotic orders for $\max_{1\le k\le L(n)} |S_{n,i}(x_k)|,\ i = 1, 2, 3.$

Lemma 4.8. *Under Assumptions 1 and 3, it holds* $\max_{1\le k\le L(n)} |S_{n,1}(x_k)| = O_P\left(\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right)$.

Proof. The proof is conducted in complete analogy to that of Lemma 4.3. We want to apply the Bernstein-type inequality for strongly mixing processes (Lemma 4.2) on $S_{n,1}(x_k)$. For this purpose, denote the truncated and centered term

$$S_{n,1}^t(x_k) = \frac{1}{nh_\sigma} \sum_{i=1}^{n} C_{K_\sigma}(X_i)\sigma^2(X_i)\left((U_i^2 - 1)1_{\{U_i^2 \le t_n\}} - \mathbb{E}\left((U_i^2 - 1)1_{\{U_i^2 \le t_n\}}\right)\right),$$

with $t_n = n^\xi$, for some $2/9 < \xi < 1/4$.

The boundedness of the ninth moments of $(U_t)_{t\in\mathbb{Z}}$ by Assumption 3 (iii) and the choice of t_n provide $P(U_n^2 > t_n) \le \frac{\mathbb{E}(|U_n|^9)}{t_n^{9/2}} = o\left(\frac{1}{n}\right)$, and

$$\mathbb{E}\left(|U_i^2 - 1|1_{\{U_i^2 > t_n\}}\right) \le \left(\mathbb{E}\left(|U_i^2 - 1|^{9/2}\right)\right)^{2/9}\left(\mathbb{E}\left(1_{\{U_i^2 > t_n\}}\right)\right)^{7/9} \le \frac{C}{t_n^{7/2}} = o\left(\frac{1}{n^{7/9}}\right),$$

since $\mathbb{E}\left(|U_i^2-1|^{9/2}\right) \leq C$ again by the moment assumption on the error process. This implies

$$\begin{aligned}\max_{1\leq k\leq L(n)} \frac{1}{nh_\sigma}\sum_{i=1}^{n} C_{K_\sigma}(X_i)\sigma^2(X_i)\mathbb{E}\left((U_i^2-1)1_{\{U_i^2\leq t_n\}}\right) &= o\left(\frac{1}{n^{7/9}h_\sigma}\right)\\ &= o\left(\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right),\end{aligned}$$

since $n^{1/2}h_\sigma \to \infty$ by Assumption 3 (vi). As in the proof of Lemma 4.3, we conclude that the truncation is asymptotically negligible.

Denote again the summands of $S_{n,1}^t(x_k)$

$$Z_{n,i} := \frac{1}{nh_\sigma}C_{K_\sigma}(X_i)\sigma^2(X_i)\left((U_i^2-1)1_{\{U_i^2\leq t_n\}} - \mathbb{E}\left((U_i^2-1)1_{\{U_i^2\leq t_n\}}\right)\right).$$

As in the proof of Lemma 4.3, the stationary process $(Z_{n,i})_{i\geq 1}$ is strongly mixing with exponentially decaying mixing coefficients $\alpha_Z(k) \leq \alpha_X(k)+\alpha_U(k)$. Thus, the Bernstein-type inequality (Lemma 4.2) with $\lambda := \eta\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}$ for some arbitrary $\eta > 0$ and $M := 4C_M t_n(nh_\sigma)^{-1} = \sup_i \|Z_{n,i}\|_\infty$ where $C_M = \sup_{x\in D}(K_\sigma(x)\sigma(x))$ provides

$$\mathrm{P}\left(|S_n^t(x_k)| \geq \eta\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right) \leq \exp\left(-\frac{C_Z\,\eta^2\frac{\log(n)}{nh_\sigma}}{v^2 n + M^2 + \eta\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}M\log(n)^2}\right),$$

where C_z is a constant that only depends on decay rate of $\alpha_Z(\cdot)$. In what follows, we show that $v^2 = O\left(\frac{1}{n^2h_\sigma}\right)$. Further,

$$\begin{aligned}M^2+\eta\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}M\log(n)^2 &= O\left(\frac{t_n^2}{(nh_\sigma)^2} + (\log(n))^{5/2}\frac{t_n}{(nh_\sigma)^{3/2}}\right)\\ &= O\left(\frac{1}{nh_\sigma}\left(\frac{1}{n^{1-2\xi}h_\sigma} + \frac{(\log(n))^{5/2}}{n^{1/2-\xi}h_\sigma^{1/2}}\right)\right)\\ &= o\left(\frac{1}{nh_\sigma}\right),\end{aligned}$$

since $\xi < 1/4$, $(\log(n))^{5/2} = O\left(n^{(1/4-\xi)/2}\right)$, and $n^{1/2}h_\sigma \to \infty$. This yields the order of $\max_{1\leq k\leq L(n)}|S_{n,1}(x_k)|$ in the manner of the proof of Lemma 4.3.

It remains to study the order of the variance term v^2. We proceed in complete analogy to the proof of Lemma 4.3. We have

$$\begin{aligned}
&\operatorname{Var}(Z_{n,1}) \\
&= \mathbb{E}\left(\frac{1}{nh_\sigma} C_{K_\sigma}(X_1)\sigma^2(X_1)\right)^2 \mathbb{E}\left((U_1^2-1)1_{\{U_1^2\le t_n\}} - \mathbb{E}\left((U_1^2-1)1_{\{U_1^2\le t_n\}}\right)\right)^2 \\
&= O\left(\frac{1}{n^2 h_\sigma}\right),
\end{aligned}$$

since the uniform boundedness of the variance function $\sigma(\cdot)$ and the assumptions on the kernel $K_\sigma(\cdot)$ provide $\mathbb{E}\left(\frac{1}{nh_\sigma} C_{K_\sigma}(X_1)\sigma^2(X_1)\right)^2 = O\left(\frac{1}{n^2 h_\sigma}\right)$, and further

$$\mathbb{E}\left((U_1^2-1)1_{\{U_1^2\le t_n\}} - \mathbb{E}\left((U_1^2-1)1_{\{U_1^2\le t_n\}}\right)\right)^2 \le \mathbb{E}\left((U_1^2+1)\right)^2 + \left(\mathbb{E}(U_1^2+1)\right)^2 \le C,$$

by the moment assumptions on the error process.

Similarly,

$$\sum_{i=2}^{\infty} |\operatorname{Cov}(Z_{n,1}, Z_{n,i})| = o\left(\frac{1}{n^2 h_\sigma}\right) + \sum_{i\ge h_\sigma^{-7/9}} |\operatorname{Cov}(Z_{n,1}, Z_{n,i})|,$$

and denoting $\widetilde{Z}_{n,i} := n\, Z_{n,i}$,

$$\left|\operatorname{Cov}\left(\widetilde{Z}_{n,1}, \widetilde{Z}_{n,i}\right)\right| \le 4(\alpha_Z(i-1))^{5/9} \left\|\widetilde{Z}_{n,1}\right\|_{9/2}^2 .$$

This time, we get $\mathbb{E}\left(|\widetilde{Z}_{n,1}|^{9/2}\right) \le C h_\sigma^{-7/2}$, and consequently $\left\|\widetilde{Z}_{n,1}\right\|_{9/2}^2 \le C h_\sigma^{-14/9}$. Proceeding as in the proof of Lemma 4.3, we get $\sum_{i\ge h_\sigma^{-7/9}} |\operatorname{Cov}(Z_{n,1}, Z_{n,i})| = o\left(\frac{1}{n^2 h_\sigma}\right)$. Consequently, $v^2 = O\left(\frac{1}{n^2 h_\sigma}\right)$. □

Lemma 4.9. *It holds* $\max_{1\le k\le L(n)} |S_{n,2}(x_k)| = O_P\left(h_0^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right)$ *under Assumption 1 and Assumption 3.*

Proof. It holds

$$\begin{aligned}
\max_{1\le k\le L(n)} |S_{n,2}(x_k)| &\le \sup_{x\in D} |m(x) - \widehat{m}(x)| \max_{1\le k\le L(n)} \frac{1}{nh_\sigma} \sum_{i=1}^{n} |C_{K_\sigma}(X_i)\, 2\sigma(X_i)|\, |U_i| \\
&=: \sup_{x\in D} |m(x) - \widehat{m}(x)| \max_{1\le k\le L(n)} \widetilde{S}_{n,2}(x_k).
\end{aligned}$$

We show that $\max_{1\le k\le L(n)} \widetilde{S}_{n,2}(x_k) = O_P(1)$. Lemma 4.3 provides $\sup_{x\in D}(m(x) - \widehat{m}(x))$

$= O_P\left(h_0^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right)$, which yields the assertion.

Denote the truncated and centered sum

$$\begin{aligned}&\widetilde{S}^t_{n,2}(x_k)\\&:= \sum_{i=1}^n \frac{1}{nh_\sigma}\left|C_{K_\sigma}(X_i)\, 2\sigma(X_i)\right| |U_i| 1_{|U_i|\le t_n} - \mathbb{E}\left|\frac{1}{h_\sigma} C_{K_\sigma}(X_1)\, 2\sigma(X_1)\right| \mathbb{E}\left(|U_1| 1_{|U_1|\le t_n}\right),\end{aligned}$$

with $t_n = n^{1/8}$. This sum term can be treated in complete analogy to $S_n^t(x_k)$ in the proof of Lemma 4.3, and hence $\max_{1\le k\le L(n)} |\widetilde{S}^t_{n,2}(x_k)| = O_P\left(\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right)$.

Note that $\widetilde{S}_{n,2}(x_k)$ is not centered, and thus centering has the cost $O(1)$. In particular, it holds

$$\begin{aligned}&|\widetilde{S}_{n,2}(x_k) - \widetilde{S}^t_{n,2}(x_k)|\\&= \frac{1}{n}\sum_{i=1}^n \left|C_{K_\sigma}(X_i)\, 2\sigma(X_i)\right| \left(|U_i| 1_{|U_i|>t_n}\right) + \mathbb{E}\left|\frac{1}{h_\sigma} C_{K_\sigma}(X_1)\, 2\sigma(X_1)\right| \mathbb{E}\left(|U_1| 1_{|U_1|\le t_n}\right)\\&=: \widetilde{I}_1(x_k) + \widetilde{I}_2(x_k),\end{aligned}$$

where $\mathrm{P}\left(\max_{1\le k\le L(n)} \widetilde{I}_1(x_k) \ne 0\right) \underset{n\to\infty}{\longrightarrow} 0$ similar to the calculations in Lemma 4.3, and the second term having the order $O(1)$: we have $\max_{1\le k\le L(n)} \mathbb{E}\left|C_{K_\sigma}(X_1)\, 2\sigma(X_1)\right| = O(1)$, since $\sigma(\cdot)$ and $f_X(\cdot)$ are uniformly bounded, and the kernel $K_\sigma(\cdot)$ is bounded with compact support. Additionally, $\mathbb{E}\left(|U_1| 1_{|U_1|\le t_n}\right) \le \sigma_U^2$. Thus, $\max_{1\le k\le L(n)} \widetilde{I}_1(x_k) = O(1)$. Consequently, $\max_{1\le k\le L(n)} \left|\widetilde{S}_{n,2}(x_k)\right| = O_P(1)$. □

Lemma 4.10. *It holds* $\max_{1\le k\le L(n)} |S_{n,3}(x_k)| = O_P\left(h_0^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right)^2$ *under Assumption 1 and Assumption 3.*

Proof. We have

$$\max_{1\le k\le L(n)} |S_{n,3}(x_k)| \le \sup_{x\in D}(m(x) - \widehat{m}(x))^2 \max_{1\le k\le L(n)} \frac{1}{nh_\sigma}\sum_{i=1}^n \left|C_{K_\sigma}(X_i)\right|.$$

Further,

$$\begin{aligned}&\max_{1\le k\le L(n)} \frac{1}{nh_\sigma}\sum_{i=1}^n \left|C_{K_\sigma}(X_i)\right|\\&= \max_{1\le k\le L(n)}\left(\frac{1}{nh_\sigma}\sum_{i=1}^n \left|C_{K_\sigma}(X_i)\right| - \mathbb{E}\left|\frac{1}{h_\sigma} C_{K_\sigma}(X_1)\right|\right) + O(1),\end{aligned}$$

since $\max\limits_{1\leq k\leq L(n)} \mathbb{E}\left|\frac{1}{h_\sigma} C_{K_\sigma}(X_1)\right| = O(1)$.

Applying Bernstein's inequality on $\frac{1}{nh_\sigma}\sum\limits_{i=1}^{n} |C_{K_\sigma}(X_i)| - \mathbb{E}\left|\frac{1}{h_\sigma} C_{K_\sigma}(X_1)\right|$ in the same manner as for $S_n^t(x_k)$ in the proof of Lemma 4.3, we get

$$\max_{1\leq k\leq L(n)} \frac{1}{nh_\sigma}\sum_{i=1}^{n} |C_{K_\sigma}(X_i)| = O_P\left(\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right) + O(1) = O_P(1).$$

Now since $\sup\limits_{x\in D}(m(x) - \widehat{m}(x))^2 = O_P\left(h_0^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right)^2$ by virtue of Lemma 4.3, we obtain the desired asymptotic order. □

Corollary 4.11. *Let D be a compact interval. Then under Assumptions 1 and 3 it holds*

$$\sup_{x\in D} |\widehat{\sigma}(x) - \sigma(x)| = O_P\left(h_\sigma^{p+1}\right) + O_P\left(\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right).$$

4.4 Efficiency of the proposed heteroscedastic estimator

In this section, we show that the feasible estimator $\widetilde{m}^{\text{het}}(x)$ is asymptotically more efficient than the conventional estimator $\widehat{m}(x)$ by demonstrating the following convergence

$$\sqrt{nh}\left(\widetilde{m}^{\text{het}}(x) - m(x) - h_n^{p+1}B(x)\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma^2_{\overline{U}^{\text{het}}}\sigma^2(x)V(x)\right).$$

The proof is conducted similar to the one in the homoscedastic case: from Corollary 2.23 we know that the infeasible estimator is more efficient than the conventional one. It thus suffices to show that the infeasible estimator $\overline{m}^{\text{het}}(x)$ and its feasible version $\widetilde{m}^{\text{het}}(x)$ are asymptotically equivalent. This is the main result of this chapter and is stated in the following Theorem:

Theorem 4.12. *Let D be a compact interval. Then under Assumptions 1 and 3 it holds*

$$\widetilde{m}^{het}(x) = \overline{m}^{het}(x) + o_P\left(\frac{1}{\sqrt{nh}}\right).$$

for all $x \in D$.

The asymptotic normality of $\widetilde{m}^{\text{het}}(x)$ is a direct consequence of Theorem 4.12:

Corollary 4.13. *Under Assumptions 1 and 3 it holds*

$$\sqrt{nh}\left(\widetilde{m}^{het}(x) - m(x) - h_n^{p+1}B(x)\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma^2_{\overline{U}^{het}}\sigma^2(x)V(x)\right),$$

for all $x \in D$, with D a compact interval.

Proof. Theorem 4.12 provides

$$\sqrt{nh}\left(\widetilde{m}^{\text{het}}(x) - m(x) - h^{p+1}B(x)\right) = \sqrt{nh}\left(\overline{m}^{\text{het}}(x) - m(x) - h^{p+1}B(x)\right) + o_P(1).$$

The assertion thus follows using Slutsky's theorem in the same manner as in Corollary 3.5 for the homoscedastic case. □

Combining the result, we thus have

$$\boxed{\begin{gathered}\sqrt{nh}\left(\widehat{m}(x) - m(x) - h^{p+1}B(x)\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma_U^2\sigma^2(x)V(x)\right),\\ \sqrt{nh}\left(\widetilde{m}^{\text{het}}(x) - m(x) - h^{p+1}B(x)\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma_{\overline{U}^{\text{het}}}^2\sigma^2(x)V(x)\right).\end{gathered}}$$

Thus, we obtain that even in the heteroscedastic case our proposed estimator is more efficient than the conventional one and should hence be preferred.

It remains to prove Theorem 4.12.

Proof of Theorem 4.12. It holds

$$\begin{aligned}\widetilde{Y}_t^{\text{het}} &= Y_t - \widehat{\sigma}(X_t)\sum_{k=1}^q \widehat{\alpha}_k\widehat{U}_{t-k}\\ &= Y_t - \sigma(X_t)\sum_{k=1}^q \alpha_k U_{t-k} - \sigma(X_t)\sum_{k=1}^q(\widehat{\alpha}_k - \alpha_k)U_{t-k} - \sigma(X_t)\sum_{k=1}^q \alpha_k(\widehat{U}_{t-k} - U_{t-k})\\ &\quad - (\widehat{\sigma}(X_t) - \sigma(X_t))\sum_{k=1}^q \alpha_k U_{t-k} - (\widehat{\sigma}(X_t) - \sigma(X_t))\sum_{k=1}^q(\widehat{\alpha}_k - \alpha_k)U_{t-k}\\ &\quad - \sigma(X_t)\sum_{k=1}^q(\widehat{\alpha}_k - \alpha_k)(\widehat{U}_{t-k} - U_{t-k}) - (\widehat{\sigma}(X_t) - \sigma(X_t))\sum_{k=1}^q \alpha_k(\widehat{U}_{t-k} - U_{t-k})\\ &\quad - (\widehat{\sigma}(X_t) - \sigma(X_t))\sum_{k=1}^q(\widehat{\alpha}_k - \alpha_k)(\widehat{U}_{t-k} - U_{t-k}).\end{aligned}$$

We thus obtain the following decomposition of $\mathbf{X}^\top\mathbf{W}\widetilde{\mathbf{Y}}^{\text{het}}$:

$$\mathbf{X}^\top\mathbf{W}\widetilde{\mathbf{Y}}^{\text{het}} = \mathbf{X}^\top\mathbf{W}\overline{\mathbf{Y}}^{\text{het}} - \mathbf{R}_{n,1} - \mathbf{R}_{n,2} - \mathbf{R}_{n,3} - \mathbf{R}_{n,4} - \mathbf{R}_{n,5} - \mathbf{R}_{n,6} - \mathbf{R}_{n,7},$$

with $\mathbf{R}_{n,i},\ i = 1, \ldots, 7$ as follows:

$$\begin{aligned}[\mathbf{R}_{n,1}]_r &:= \frac{1}{nh}\sum_{t=1}^n K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1}\sigma(X_t)\sum_{k=1}^q(\widehat{\alpha}_k - \alpha_k)U_{t-k} = o_P\left(\frac{1}{nh}\right),\\ [\mathbf{R}_{n,2}]_r &:= \frac{1}{nh}\sum_{t=1}^n K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1}\sigma(X_t)\sum_{k=1}^q \alpha_k(\widehat{U}_{t-k} - U_{t-k})\end{aligned}$$

$$
\begin{aligned}
&= O_P\left(h_0^{p+1} + h_\sigma^{p+1} + \frac{\log(n)}{n\sqrt{h_0 h_\sigma}} + \frac{\log(n)}{nh_\sigma}\right) + o_P\left(\frac{1}{\sqrt{nh}}\right), \\
[\mathbf{R}_{n,3}]_r &:= \frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1} (\widehat{\sigma}(X_t) - \sigma(X_t)) \sum_{k=1}^{q} \alpha_k U_{t-k} \\
&= O_P(h_\sigma^{p+1}) + o_P\left(\frac{1}{\sqrt{nh}}\right), \\
[\mathbf{R}_{n,4}]_r &:= \frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1} (\widehat{\sigma}(X_t) - \sigma(X_t)) \sum_{k=1}^{q} (\widehat{\alpha}_k - \alpha_k) U_{t-k} \\
&= o_P\left(\frac{1}{\sqrt{nh}}\left(h_\sigma^{p+1} + \frac{1}{\sqrt{nh}}\right)\right), \\
[\mathbf{R}_{n,5}]_r &:= \frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1} \sigma(X_t) \sum_{k=1}^{q} (\widehat{\alpha}_k - \alpha_k)(\widehat{U}_{t-k} - U_{t-k}) \\
&= o_P\left(\frac{1}{\sqrt{nh}}\left(h_0^{p+1} + h_\sigma^{p+1} + \frac{\log(n)}{n\sqrt{h_0 h_\sigma}} + \frac{\log(n)}{nh_\sigma} + \frac{1}{\sqrt{nh}}\right)\right), \\
[\mathbf{R}_{n,6}]_r &:= \frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1} (\widehat{\sigma}(X_t) - \sigma(X_t)) \sum_{k=1}^{q} \alpha_k (\widehat{U}_{t-k} - U_{t-k}) \\
&= O_P\left(h_0^{p+1} + h_\sigma^{p+1} + \frac{\log(n)}{nh_\sigma} + \frac{\log(n)}{n\sqrt{h_0 h_\sigma}}\right), \\
[\mathbf{R}_{n,7}]_r &:= \frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1} (\widehat{\sigma}(X_t) - \sigma(X_t)) \\
&\quad \times \sum_{k=1}^{q} (\widehat{\alpha}_k - \alpha_k)(\widehat{U}_{t-k} - U_{t-k}) \\
&= o_P\left(\frac{1}{\sqrt{nh}}\left(h_0^{p+1} + h_\sigma^{p+1} + \frac{\log(n)}{nh_\sigma} + \frac{\log(n)}{n\sqrt{h_0 h_\sigma}}\right)\right),
\end{aligned}
$$

where the orders of convergence will be verified in Lemmas 4.21 - 4.27.

We comment on the asymptotic rates. Decomposition of the residuals gives

$$
U_t - \widehat{U}_t = \left(\frac{\widehat{\sigma}(X_t) - \sigma(X_t)}{\widehat{\sigma}(X_t)}\right) U_t + \frac{\widehat{m}(X_t) - m(X_t)}{\widehat{\sigma}(X_t)}.
$$

This decomposition implies that the asymptotic order of $\mathbf{R}_{n,2}$ is determined by that of the variance estimation and the prior local polynomial smoothing both averaged by the kernel function. We already studied the latter in the last chapter, and the former can be made asymptotically negligible by a suitable choice of h_σ. The estimation of $\boldsymbol{\alpha}$ also depends on the estimation of the error process, and thus - by a suitable choice of the prior bandwidths - we obtain the convergence rate $o_P\left(\frac{1}{\sqrt{nh}}\right)$ for the filtering parameter $\boldsymbol{\alpha}$ (see Lemma 4.17 and Assumption 3 (vi)). This implies the rate of $\mathbf{R}_{n,1}$ (note that we studied the remaining term in the previous chapter). The convergence rates of the remaining terms follow as a direct combination of the first two rates and the uniform convergence results provided in the previous section.

We thus have

$$\mathbf{R}_{n,i} = O_P\left(h_0^{p+1} + h_\sigma^{p+1} + \frac{\log(n)}{n\sqrt{h_0 h_\sigma}} + \frac{\log(n)}{nh_\sigma}\right) + o_P\left(\frac{1}{\sqrt{nh}}\right), \ \ i = 1, \ldots, 7.$$

Assumption 3 (vi) on the bandwidths implies $\mathbf{R}_{n,i} = o_P\left(\frac{1}{\sqrt{nh}}\right)$, $i = 1, \ldots, 7$. Substituting these expressions into $\widetilde{m}^{\text{het}}(x)$, we have $\widetilde{m}^{\text{het}}(x) = \overline{m}^{\text{het}}(x) - \sum_{i=1}^{7} \mathbf{Q}_{n,i}$, where $\mathbf{Q}_{n,i} := \mathbf{e}_1^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{R}_{n,i} = o_P\left(\frac{1}{\sqrt{nh}}\right)$, $i = 1, \ldots, 7$ similar to former calculations. This yields the assertion. □

It remains to establish the asymptotic orders of the estimator $\widehat{\boldsymbol{\alpha}}$ for the filtering parameter, and those of the rest terms $\mathbf{R}_{n,i}, \ i = 1, \ldots, 7$. We begin with the asymptotic order for the filtering parameter $\boldsymbol{\alpha}$. We will need the following auxiliary Lemmas that give upper bounds for the mean of the summed product of vector-valued strongly mixing processes:

Lemma 4.14. *Let $\left(Z_t^{(1)}, \ldots, Z_t^{(6)}\right)_{t\geq 1}^\mathsf{T}$ be a strongly mixing process with exponentially decreasing strong mixing coefficients $\alpha_Z(\cdot)$. Further let $\mathbb{E}\left(Z_{t_i}^{(i)}\right) = 0$ for all $1 \leq t_i \leq n$ and $i = 1, \ldots, 6$, and as well $\left\|Z_{t_1}^{(1)}\right\|_{k_1} \cdots \left\|Z_{t_6}^{(6)}\right\|_{k_6} \leq C$, for some $\frac{1}{k_1} + \ldots + \frac{1}{k_6} = 1 - C_\alpha$, with $C_\alpha \in (0,1)$. Then it holds*

$$\sum_{1\leq t_1 \leq \ldots \leq t_6 \leq n} \left|\mathbb{E}\left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)}\right)\right| = O\left(n^3\right).$$

Proof. We show that

$$\begin{aligned}
&\left|\mathbb{E}\left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)}\right)\right| \\
&\leq C\,\alpha_Z\left(\max_{1\leq i\leq 5}\{t_{i+1} - t_i\}\right)^{C_\alpha} + \left|\mathbb{E}\left(Z_{t_1}^{(1)} Z_{t_2}^{(2)}\right)\right| \left|\mathbb{E}\left(Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)}\right)\right| \\
&\quad + \left|\mathbb{E}\left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)}\right)\right| \left|\mathbb{E}\left(Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)}\right)\right| + \left|\mathbb{E}\left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} Z_{t_4}^{(4)}\right)\right| \left|\mathbb{E}\left(Z_{t_5}^{(5)} Z_{t_6}^{(6)}\right)\right|,
\end{aligned} \tag{4.8}$$

distinguishing the different cases for the maximum distance $T := \max_{1\leq i\leq 5}\{t_{i+1} - t_i\}$.

(i) $T = t_2 - t_1$.
The covariance inequality for strongly mixing processes (Lemma 2.5) provides

$$\begin{aligned}
\left|\mathbb{E}\left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)}\right)\right| &= \left|\mathrm{Cov}\left(Z_{t_1}^{(1)}, Z_{t_2}^{(2)} Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)}\right)\right| \\
&\leq 4\left(\alpha_Z(T)\right)^{C_\alpha} \left\|Z_{t_1}^{(1)}\right\|_{k_1} \left\|Z_{t_2}^{(2)} Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)}\right\|_{k},
\end{aligned}$$

with $\frac{1}{k} := \frac{1}{k_2} + \ldots + \frac{1}{k_6}$.

Hölder's inequality gives

$$\left\| Z_{t_2}^{(2)} Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)} \right\|_k \leq \prod_{i=2}^{6} \left\| Z_{t_i}^{(i)} \right\|_{k_i},$$

and thus

$$\left| \mathbb{E} \left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)} \right) \right| \leq C \left(\alpha_Z(T) \right)^{C_\alpha}.$$

(ii) $T = t_3 - t_2$.
It holds

$$\begin{aligned} &\left| \mathbb{E} \left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)} \right) \right| \\ &\leq \left| \mathrm{Cov} \left(Z_{t_1}^{(1)} Z_{t_2}^{(2)}, Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)} \right) \right| + \left| \mathbb{E} \left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} \right) \right| \left| \mathbb{E} \left(Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)} \right) \right|. \end{aligned}$$

Another usage of the covariance inequality for strongly mixing processes gives

$$\left| \mathrm{Cov} \left(Z_{t_1}^{(1)} Z_{t_2}^{(2)}, Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)} \right) \right| \leq 4 \left(\alpha_Z(T) \right)^{C_\alpha} \left\| Z_{t_1}^{(1)} Z_{t_2}^{(2)} \right\|_k \left\| Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)} \right\|_{\widetilde{k}},$$

with $\frac{1}{k} := \frac{1}{k_1} + \frac{1}{k_2}$, and $\frac{1}{\widetilde{k}} := \frac{1}{k_3} + \frac{1}{k_4} + \frac{1}{k_5} + \frac{1}{k_6}$. Applying Hölder's inequality on the two product terms yields the assertion.

(iii) $T = t_4 - t_3$.
Similar to former calculations we obtain with $\frac{1}{k} := \frac{1}{k_1} + \frac{1}{k_2} + \frac{1}{k_3}$, and $\frac{1}{\widetilde{k}} := \frac{1}{k_4} + \frac{1}{k_5} + \frac{1}{k_6}$

$$\begin{aligned} &\left| \mathbb{E} \left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)} \right) \right| \\ &\leq \left| \mathrm{Cov} \left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)}, Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)} \right) \right| + \left| \mathbb{E} \left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} \right) \right| \left| \mathbb{E} \left(Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)} \right) \right| \\ &\leq 4 \left(\alpha_Z(T) \right)^{C_\alpha} \left\| Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} \right\|_k \left\| Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)} \right\|_{\widetilde{k}} \\ &\quad + \left| \mathbb{E} \left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} \right) \right| \left| \mathbb{E} \left(Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)} \right) \right| \\ &\leq C \left(\alpha_Z(T) \right)^{C_\alpha} + \left| \mathbb{E} \left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} \right) \right| \left| \mathbb{E} \left(Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)} \right) \right|. \end{aligned}$$

(iv) $T = t_5 - t_4$.
This case is similar to (ii), providing

$$\begin{aligned} &\left| \mathbb{E} \left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)} \right) \right| \\ &\leq C \left(\alpha_Z(T) \right)^{C_\alpha} + \left| \mathbb{E} \left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} Z_{t_4}^{(4)} \right) \right| \left| \mathbb{E} \left(Z_{t_5}^{(5)} Z_{t_6}^{(6)} \right) \right|. \end{aligned}$$

(v) $T = t_6 - t_5$.

In analogy to case (i), we get

$$\left|\mathbb{E}\left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)}\right)\right| \leq C\,(\alpha_Z(T))^{C_\alpha}.$$

Now note that

$$\begin{aligned}
&\sum_{1\leq t_1\leq\ldots\leq t_6\leq n}\left(\alpha_Z\left(\max_{1\leq i\leq 5}\{t_{i+1}-t_i\}\right)\right)^{C_\alpha}\\
&=\sum_{T=0}^{n}(\alpha_Z(T))^{C_\alpha}\,\#\left\{(1\leq t_1\leq\ldots\leq t_6\leq n):\max_{1\leq i\leq 5}\{t_{i+1}-t_i\}=T\right\}\\
&\leq\ 5\,n\sum_{T=0}^{n}(\alpha_Z(T))^{C_\alpha}(T+1)^4\\
&=C\,n=O(n),
\end{aligned}\tag{4.9}$$

since the strong mixing coefficients $\alpha_Z(\cdot)$ are exponentially decreasing, and since the number of time points $(1\leq t_1\leq t_2\leq\ldots\leq t_6\leq n)$ with maximal distance T can be calculated as follows:

$$\begin{aligned}
&\#\left\{(1\leq t_1\leq\ldots\leq t_6\leq n):\max_{1\leq i\leq 5}\{t_{i+1}-t_i\}=T\right\}\\
&=5\,\#\left\{t_2=t_1+T,\ \text{and } t_{i+1}\in\{t_i,\ldots,t_i+T\} \text{ for all } i=2,\ldots,5\right\}\\
&\leq 5\,n(T+1)^4.
\end{aligned}$$

Further, we have

$$\begin{aligned}
&\left|\mathbb{E}\left(Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)}\right)\right|\\
&\leq\left|\mathrm{Cum}\left(Z_{t_3}^{(3)}, Z_{t_4}^{(4)}, Z_{t_5}^{(5)}, Z_{t_6}^{(6)}\right)\right|+\left|\mathbb{E}\left(Z_{t_3}^{(3)} Z_{t_4}^{(4)}\right)\right|\left|\mathbb{E}\left(Z_{t_5}^{(5)} Z_{t_6}^{(6)}\right)\right|\\
&\quad+\left|\mathbb{E}\left(Z_{t_3}^{(3)} Z_{t_5}^{(5)}\right)\right|\left|\mathbb{E}\left(Z_{t_4}^{(4)} Z_{t_6}^{(6)}\right)\right|+\left|\mathbb{E}\left(Z_{t_3}^{(3)} Z_{t_5}^{(5)}\right)\right|\left|\mathbb{E}\left(Z_{t_4}^{(4)} Z_{t_6}^{(6)}\right)\right|.
\end{aligned}$$

Similar calculations as in the proof of Lemma 3.7 provide

$$\sum_{1\leq t_3\leq\ldots\leq t_6\leq n}\left|\mathrm{Cum}\left(Z_{t_3}^{(3)}, Z_{t_4}^{(4)}, Z_{t_5}^{(5)}, Z_{t_6}^{(6)}\right)\right|=O(n).$$

Further note that for $i<j\in\{1,\ldots,6\}$ arbitrary the covariance inequality for strongly mixing processes yields

$$\sum_{1\leq t_i\leq t_j\leq n}\left|\mathbb{E}\left(Z_{t_i}^{(i)} Z_{t_j}^{(j)}\right)\right|\leq 4\sum_{1\leq t_i\leq t_j\leq n}(\alpha_Z(t_j-t_i))^{C_\alpha}\left\|Z_{t_i}^{(i)}\right\|_{k_i}\left\|Z_{t_j}^{(j)}\right\|_{k_j}=O(n),$$

again since the mixing coefficients are exponentially decreasing.

Hence $\sum_{t_3\leq\ldots\leq t_6}|\mathbb{E}(Z_{t_3}^{(3)}Z_{t_4}^{(4)}Z_{t_5}^{(5)}Z_{t_6}^{(6)})|=O(n^2)$, and thus

$$\sum_{1\leq t_1\leq t_2\leq n}\left|\mathbb{E}\left(Z_{t_1}^{(1)}Z_{t_2}^{(2)}\right)\right|\sum_{1\leq t_3\leq\ldots\leq t_6\leq n}\left|\mathbb{E}\left(Z_{t_3}^{(3)}Z_{t_4}^{(4)}Z_{t_5}^{(5)}Z_{t_6}^{(6)}\right)\right|=O(n)O\left(n^2\right)=O\left(n^3\right), \tag{4.10}$$

and similarly

$$\sum_{1\leq t_1\leq\ldots\leq t_6\leq n}\left|\mathbb{E}\left(Z_{t_1}^{(1)}Z_{t_2}^{(2)}Z_{t_3}^{(3)}Z_{t_4}^{(4)}\right)\right|\left|\mathbb{E}\left(Z_{t_5}^{(5)}Z_{t_6}^{(6)}\right)\right|=O\left(n^3\right). \tag{4.11}$$

For the remaining sum term, the covariance inequality for strongly mixing processes provides

$$\begin{aligned}\left|\mathbb{E}\left(Z_{t_1}^{(1)}Z_{t_2}^{(2)}Z_{t_3}^{(3)}\right)\right| &= \min\left\{\left|\operatorname{Cov}\left(Z_{t_1}^{(1)},Z_{t_2}^{(2)}Z_{t_3}^{(3)}\right)\right|,\left|\operatorname{Cov}\left(Z_{t_1}^{(1)}Z_{t_2}^{(2)},Z_{t_3}^{(3)}\right)\right|\right\}\\ &\leq 4\ \left(\alpha_Z\left(\max\{t_2-t_1,t_3-t_2\}\right)\right)^{C_\alpha}\left\|Z_{t_1}^{(1)}\right\|_{k_1}\left\|Z_{t_2}^{(2)}\right\|_{k_2}\left\|Z_{t_3}^{(3)}\right\|_{k_3}\\ &\leq C\ (\alpha_Z(\max\{t_2-t_1,t_3-t_2\}))^{C_\alpha}.\end{aligned}$$

Now similar to equation (4.9) we get

$$\begin{aligned}\sum_{1\leq t_1\leq t_2\leq t_3\leq n}\left|\mathbb{E}\left(Z_{t_1}^{(1)}Z_{t_2}^{(2)}Z_{t_3}^{(3)}\right)\right| &\leq C\sum_{1\leq t_1\leq t_2\leq t_3\leq n}(\alpha_Z(\max\{t_2-t_1,t_3-t_2\}))^{C_\alpha}\\ &\leq C\ 2n\sum_{T=0}^{\infty}(T+1)(\alpha_Z(T))^{C_\alpha}\\ &= O(n).\end{aligned}$$

Similar calculations for $\sum_{1\leq t_4\leq t_5\leq t_6\leq n}|\mathbb{E}(Z_{t_4}^{(4)}Z_{t_5}^{(5)}Z_{t_6}^{(6)})|$ provide

$$\sum_{1\leq t_1\leq\ldots\leq t_6\leq n}\left|\mathbb{E}\left(Z_{t_1}^{(1)}Z_{t_2}^{(2)}Z_{t_3}^{(3)}\right)\right|\left|\mathbb{E}\left(Z_{t_4}^{(4)}Z_{t_5}^{(5)}Z_{t_6}^{(6)})\right)\right|=O\left(n^2\right). \tag{4.12}$$

Combining equations (4.8)-(4.12) we conclude

$$\begin{aligned}&\sum_{1\leq t_1\leq\ldots\leq t_6\leq n}\left|\mathbb{E}\left(Z_{t_1}^{(1)}Z_{t_2}^{(2)}Z_{t_3}^{(3)}Z_{t_4}^{(4)}Z_{t_5}^{(5)}Z_{t_6}^{(6)}\right)\right|\\ &\leq\sum_{1\leq t_1\leq\ldots\leq n}\left(C\left(\alpha_Z\left(\max_{1\leq i\leq 5}\{t_{i+1}-t_i\}\right)\right)^{C_\alpha}+\left|\mathbb{E}\left(Z_{t_1}^{(1)}Z_{t_2}^{(2)}\right)\right|\left|\mathbb{E}\left(Z_{t_3}^{(3)}Z_{t_4}^{(4)}Z_{t_5}^{(5)}Z_{t_6}^{(6)}\right)\right|\right.\\ &\quad\left.+\left|\mathbb{E}\left(Z_{t_1}^{(1)}Z_{t_2}^{(2)}Z_{t_3}^{(3)}\right)\right|\left|\mathbb{E}\left(Z_{t_4}^{(4)}Z_{t_5}^{(5)}Z_{t_6}^{(6)}\right)\right|+\left|\mathbb{E}\left(Z_{t_1}^{(1)}Z_{t_2}^{(2)}Z_{t_3}^{(3)}Z_{t_4}^{(4)}\right)\right|\left|\mathbb{E}\left(Z_{t_5}^{(5)}Z_{t_6}^{(6)}\right)\right|\right)\\ &=O(n)+O\left(n^3\right)+O\left(n^2\right)+O\left(n^3\right)\\ &=O\left(n^3\right),\end{aligned}$$

which yields the assertion. □

Corollary 4.15. *Under the assumptions of Lemma 4.14 we have*

$$\sum_{t_1,t_2,\ldots t_6} \left|\mathbb{E}\left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)}\right)\right| = O\left(n^3\right),$$

with $t_i = 1, \ldots, n$ *for all* $i = 1, \ldots, 6$.

Proof. It holds

$$\begin{aligned}
&\sum_{t_1,t_2,\ldots t_6} \left|\mathbb{E}\left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)}\right)\right| \\
&= 6! \sum_{1\le t_1 \le \ldots \le t_6 \le n} \left|\mathbb{E}\left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} Z_{t_4}^{(4)} Z_{t_5}^{(5)} Z_{t_6}^{(6)}\right)\right| \\
&= O\left(n^3\right).
\end{aligned}$$

□

Lemma 4.16. *Let* $\left(Z_t^{(1)}, \ldots, Z_t^{(4)}\right)^{\mathsf{T}}_{t\in\mathbb{Z}}$ *be a strongly mixing process with exponentially decreasing strong mixing coefficients* α_Z. *Further let* $Z_{t_i}^{(i)}$ *be centered for all* $i = 1, \ldots, 4$, *and* $1 \le t_i \le n$, *and let* $\left\|Z_{t_1}^{(1)}\right\|_{k_1} \cdots \left\|Z_{t_4}^{(4)}\right\|_{k_4} \le C$, *for some* $\frac{1}{k_1} + \ldots + \frac{1}{k_4} := 1 - C_\alpha$, *with* $C_\alpha \in (0,1)$. *Then it holds*

$$\sum_{t_1,t_2,t_3,t_4} \left|\mathbb{E}\left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} Z_{t_4}^{(4)}\right)\right| = O\left(n^2\right).$$

Proof. It holds

$$\sum_{t_1,t_2,t_3,t_4} \left|\mathbb{E}\left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} Z_{t_4}^{(4)}\right)\right| = 4! \sum_{1\le t_1\le t_2\le t_3\le t_4\le n} \left|\mathbb{E}\left(Z_{t_1}^{(1)} Z_{t_2}^{(2)} Z_{t_3}^{(3)} Z_{t_4}^{(4)}\right)\right| = O\left(n^2\right),$$

where the last equality follows directly from the proof of Lemma 4.14. □

With the help of this auxiliary Lemmas, we can now develop the asymptotic order of the estimation of the filtering parameter $\boldsymbol{\alpha}$:

Lemma 4.17 (Upper bound for least squares estimation of $\boldsymbol{\alpha}$)**.** *Under Assumptions 1 and 3, it holds*

$$\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\| = O_P\left(h_0^{p+1} + h_\sigma^{p+1} + \frac{1}{n\sqrt{hh_\sigma}} + \frac{\log(n)}{n}\left(\frac{1}{h_0} + \frac{1}{\sqrt{h_\sigma h_0}} + \frac{1}{h_\sigma}\right) + \frac{1}{\sqrt{n}}\right).$$

Proof. The proof is conducted similar to that in the homoscedastic case (see Lemma 3.9). We decompose

$$\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} = \widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}} + \widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha},$$

In total analogy we obtain $\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}} = O_P\left(n^{-1/2}\right)$ for the first term.

For the second term, denote again

$$\mathbf{G} := \frac{1}{n}\mathbf{U}_q^\intercal \mathbf{U}_q,\ \widehat{\mathbf{G}} := \frac{1}{n}\widehat{\mathbf{U}}_q^\intercal \widehat{\mathbf{U}}_q,\ \text{and } \mathbf{g} := \frac{1}{n}\mathbf{U}_q^\intercal \mathbf{U},\ \widehat{\mathbf{g}} := \frac{1}{n}\widehat{\mathbf{U}}_q^\intercal \widehat{\mathbf{U}},$$

to get

$$\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}} = \mathbf{G}^{-1}(\widehat{\mathbf{g}} - \mathbf{g}) - \widehat{\mathbf{G}}^{-1}(\widehat{\mathbf{G}} - \mathbf{G})\mathbf{G}^{-1}\mathbf{g} - \widehat{\mathbf{G}}^{-1}(\widehat{\mathbf{G}} - \mathbf{G})\mathbf{G}^{-1}(\widehat{\mathbf{g}} - \mathbf{g}).$$

We derive the following convergence rates:

$$\begin{aligned} \widehat{\mathbf{G}} - \mathbf{G} &= O_P\left(h_0^{p+1} + h_\sigma^{p+1} + \frac{1}{n\sqrt{hh_\sigma}} + \frac{\log(n)}{n}\left(\frac{1}{h_0} + \frac{1}{\sqrt{h_\sigma h_0}} + \frac{1}{h_\sigma}\right) + \frac{1}{\sqrt{n}}\right), \\ \widehat{\mathbf{g}} - \mathbf{g} &= O_P\left(h_0^{p+1} + h_\sigma^{p+1} + \frac{1}{n\sqrt{hh_\sigma}} + \frac{\log(n)}{n}\left(\frac{1}{h_0} + \frac{1}{\sqrt{h_\sigma h_0}} + \frac{1}{h_\sigma}\right) + \frac{1}{\sqrt{n}}\right). \end{aligned} \tag{4.13}$$

As in the homoscedastic case, this directly implies $\mathbf{G}^{-1} = O_P(1)$, $\widehat{\mathbf{G}}^{-1} = O_P(1)$, and $\mathbf{g} = O_P(1)$. We therefore obtain

$$\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}} = O_P\left(h_0^{p+1} + h_\sigma^{p+1} + \frac{1}{n\sqrt{hh_\sigma}} + \frac{\log(n)}{n}\left(\frac{1}{h_0} + \frac{1}{\sqrt{h_\sigma h_0}} + \frac{1}{h_\sigma}\right) + \frac{1}{\sqrt{n}}\right),$$

which yields the assertion.

We proceed element-wise. We have $[\widehat{\mathbf{G}}]_{k,l} = \frac{1}{n}\sum\limits_{t=1}^{n} \widehat{U}_{t-k}\widehat{U}_{t-l}$, and

$$\widehat{U}_t = \frac{B_{X_t} + V_{X_t}}{\widehat{\sigma}(X_t)} + \frac{\sigma(X_t)}{\widehat{\sigma}(X_t)}U_t.$$

Hence

$$\begin{aligned} &\widehat{U}_{t-k}\widehat{U}_{t-l} - U_{t-k}U_{t-l} \\ &= \left(\frac{B_{X_{t-k}} + V_{X_{t-k}}}{\widehat{\sigma}(X_{t-k})} + \frac{\sigma(X_{t-k})}{\widehat{\sigma}(X_{t-k})}U_{t-k}\right)\left(\frac{B_{X_{t-l}} + V_{X_{t-l}}}{\widehat{\sigma}(X_{t-l})} + \frac{\sigma(X_{t-l})}{\widehat{\sigma}(X_{t-l})}U_{t-l}\right) - U_{t-k}U_{t-l} \\ &= \frac{1}{\widehat{\sigma}(X_{t-k})\widehat{\sigma}(X_{t-l})}\big((B_{X_{t-k}} + V_{X_{t-k}})(B_{X_{t-l}} + V_{X_{t-l}}) + (B_{X_{t-k}} + V_{X_{t-k}})U_{t-l}\sigma(X_{t-l}) \\ &\quad + (B_{X_{t-l}} + V_{X_{t-l}})U_{t-k}\sigma(X_{t-k})\big) + \left(\frac{\sigma(X_{t-k})\sigma(X_{t-l})}{\widehat{\sigma}(X_{t-k})\widehat{\sigma}(X_{t-l})} - 1\right)U_{t-k}U_{t-l}. \end{aligned}$$

We decompose the term $1/\widehat{\sigma}(\cdot)$:

$$\begin{aligned}\frac{1}{\widehat{\sigma}(x)\widehat{\sigma}(y)} &= \frac{1}{\sigma(x)\sigma(y)} + \frac{\sigma(x)-\widehat{\sigma}(x)}{\sigma(x)\widehat{\sigma}(x)\sigma(y)} + \frac{\sigma(y)-\widehat{\sigma}(y)}{\sigma(y)\widehat{\sigma}(y)\sigma(x)} + \frac{(\sigma(x)-\widehat{\sigma}(x))(\sigma(y)-\widehat{\sigma}(y))}{\sigma(x)\sigma(y)\widehat{\sigma}(x)\widehat{\sigma}(y)} \\ &=: \frac{1}{\sigma(x)\sigma(y)} + M_{\sigma,\widehat{\sigma}}(x,y).\end{aligned}$$

Therefore, denoting $B_{X_{t-k}} + V_{X_{t-k}} =: R_B^V(k)$ for notational convenience, it holds

$$\begin{aligned}&\frac{1}{n}\sum_{t=1}^{n}(\widehat{U}_{t-k}\widehat{U}_{t-l} - U_{t-k}U_{t-l}) \\ &= \frac{1}{n}\sum_{t=1}^{n}\frac{1}{\widehat{\sigma}(X_{t-k})\widehat{\sigma}(X_{t-l})}\left(R_B^V(k)R_B^V(l) + R_B^V(k)U_{t-l}\sigma(X_{t-l}) + R_B^V(l)U_{t-k}\sigma(X_{t-k})\right) \\ &\quad + \frac{1}{n}\sum_{t=1}^{n}\left(\frac{\sigma(X_{t-k})\sigma(X_{t-l})}{\widehat{\sigma}(X_{t-k})\widehat{\sigma}(X_{t-l})} - 1\right)U_{t-k}U_{t-l} \\ &= \frac{1}{n}\sum_{t=1}^{n}\frac{1}{\sigma(X_{t-k})\sigma(X_{t-l})}\left(R_B^V(k)R_B^V(l) + R_B^V(k)U_{t-l}\sigma(X_{t-l}) + R_B^V(l)U_{t-k}\sigma(X_{t-k})\right) \\ &\quad + \frac{1}{n}\sum_{t=1}^{n}M_{\sigma,\widehat{\sigma}}(X_{t-k},X_{t-l})\left(R_B^V(k)R_B^V(l) + R_B^V(k)U_{t-l}\sigma(X_{t-l}) + R_B^V(l)U_{t-k}\sigma(X_{t-k})\right) \\ &\quad + \frac{1}{n}\sum_{t=1}^{n}\left(\frac{\sigma(X_{t-k})\sigma(X_{t-l})}{\widehat{\sigma}(X_{t-k})\widehat{\sigma}(X_{t-l})} - 1\right)U_{t-k}U_{t-l} \\ &=: S_{n,1} + S_{n,2} + S_{n,3}.\end{aligned}$$

We show that

$$\begin{aligned}S_{n,1} &= O_P\left(h_0^{p+1}\right) + O_P\left(\frac{1}{nh_0}\right), \\ S_{n,2} &= O_P\left(\left(h_\sigma^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right)\left(h_0^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right)\right), \\ S_{n,3} &= O_P\left(\left(h_\sigma^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right)^2 + h_\sigma^{p+1} + h_0^{p+1} + \frac{\log(n)}{nh_0} + \frac{1}{\sqrt{n}}\right).\end{aligned}$$

The rate of $S_{n,1}$ follows directly from the calculations for $\widetilde{\boldsymbol{\alpha}}$ in the homoscedastic case, and the asymptotic order of $S_{n,2}$ is a consequence of the uniform convergence results for both the local polynomial and the variance estimators. The last term $S_{n,3}$ requires a more detailed analysis. These rates imply

$$\widehat{\mathbf{G}} - \mathbf{G} = O_P\left(h_0^{p+1} + h_\sigma^{p+1} + \frac{1}{n\sqrt{hh_\sigma}} + \frac{\log(n)}{n}\left(\frac{1}{h_0} + \frac{1}{\sqrt{h_\sigma h_0}} + \frac{1}{h_\sigma}\right) + \frac{1}{\sqrt{n}}\right).$$

Similar calculations provide

$$\widehat{\mathbf{g}} - \mathbf{g} = O_P\left(h_0^{p+1} + h_\sigma^{p+1} + \frac{1}{n\sqrt{hh_\sigma}} + \frac{\log(n)}{n}\left(\frac{1}{h_0} + \frac{1}{\sqrt{h_\sigma h_0}} + \frac{1}{h_\sigma}\right) + \frac{1}{\sqrt{n}}\right),$$

which concludes the proof.

Asymptotic order of $S_{n,1}$. Since the variance function $\sigma(\cdot)$ is both bounded and bounded away from zero, in analogy to the proof of Lemma 3.9 we obtain the same convergence rates as stated therein:

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{\sigma(X_{t-k})\sigma(X_{t-l})}\right)\big((B_{X_{t-k}} + V_{X_{t-k}})(B_{X_{t-l}} + V_{X_{t-l}}) \\
&\quad + (B_{X_{t-k}} + V_{X_{t-k}})U_{t-l}\sigma(X_{t-l}) + (B_{X_{t-l}} + V_{X_{t-l}})U_{t-k}\sigma(X_{t-k})\big) \\
&= O_P\left(h_0^{p+1}\right) + O_P\left(\frac{1}{nh_0}\right).
\end{aligned}$$

Asymptotic order of $S_{n,2}$. We have

$$\begin{aligned}
|S_{n,2}| &= \Bigg|\frac{1}{n}\sum_{t=1}^{n} M_{\sigma,\widehat{\sigma}}(X_{t-k}, X_{t-l})\big((B_{X_{t-k}} + V_{X_{t-k}})(B_{X_{t-l}} + V_{X_{t-l}}) \\
&\quad + (B_{X_{t-k}} + V_{X_{t-k}})U_{t-l}\sigma(X_{t-l}) + (B_{X_{t-l}} + V_{X_{t-l}})U_{t-k}\sigma(X_{t-k})\big)\Bigg| \\
&\le \sup_{\{x,y\}\in D}|M_{\sigma,\widehat{\sigma}}(x,y)|\,\frac{1}{n}\sum_{t=1}^{n}\big|(B_{X_{t-k}} + V_{X_{t-k}})(B_{X_{t-l}} + V_{X_{t-l}})\big| \\
&\quad + \sup_{\{x,y\}\in D}|M_{\sigma,\widehat{\sigma}}(x,y)|C_{\sigma,2}\,\frac{1}{n}\sum_{t=1}^{n}\big|(B_{X_{t-k}} + V_{X_{t-k}})U_{t-l} + (B_{X_{t-l}} + V_{X_{t-l}})U_{t-k}\big|.
\end{aligned}$$

It holds

$$\begin{aligned}
&\sup_{\{x,y\}\in D}|M_{\sigma,\widehat{\sigma}}(x,y)| \\
&= \sup_{\{x,y\}\in D}\left|\frac{\sigma(x)-\widehat{\sigma}(x)}{\sigma(x)\widehat{\sigma}(x)\sigma(y)} + \frac{\sigma(y)-\widehat{\sigma}(y)}{\sigma(y)\widehat{\sigma}(y)\sigma(x)} + \frac{(\sigma(x)-\widehat{\sigma}(x))(\sigma(y)-\widehat{\sigma}(y))}{\sigma(x)\sigma(y)\widehat{\sigma}(x)\widehat{\sigma}(y)}\right| \\
&\le \left(\left(\sup_{z\in D}|\sigma(z)-\widehat{\sigma}(z)|\right)\left(\sup_{z\in D}\left|\frac{1}{\widehat{\sigma}(z)}\right|\right)\left(\sup_{z\in D}\left|\frac{1}{\sigma(z)}\right|\right)^2\right) \\
&\quad \times\left(2 + \left(\sup_{z\in D}|\sigma(z)-\widehat{\sigma}(z)|\right)\left(\sup_{z\in D}\left|\frac{1}{\widehat{\sigma}(z)}\right|\right)\right).
\end{aligned}$$

Now note that $\sigma(\cdot) \ge C_{\sigma,1}$, and $\sup_{z\in D}|\sigma(z)-\widehat{\sigma}(z)| = O_P\left(h_\sigma^{p+1}\right) + O_P\left(\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right)$ from the uniform convergence results of $\sigma(\cdot)$ by Lemma 4.7. Further, again by virtue of the uniform convergence rate for $\sup_{z\in D}|\sigma(z)-\widehat{\sigma}(z)|$ and the boundedness of $1/\sigma(\cdot)$ we get

$$\inf_{z\in D}\widehat{\sigma}(z)\geq\inf_{z\in D}\sigma(z)-\sup_{z\in D}(|\sigma(z)-\widehat{\sigma}(z)|)\geq C_{\sigma,1}-\left(O_P\left(h_\sigma^{p+1}\right)+O_P\left(\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right)\right).$$

Therefore we have for $0<\varepsilon<C_{\sigma,1}$ arbitrary

$$\mathrm{P}\left(\sup_{z\in D}\left(\frac{1}{\widehat{\sigma}(z)}\right)>\frac{1}{C_{\sigma,1}-\varepsilon}\right)=\mathrm{P}\left(\inf_{z\in D}\widehat{\sigma}(z)<C_{\sigma,1}-\varepsilon\right)\underset{n\to\infty}{\longrightarrow}0,$$

and hence

$$\sup_{z\in D}\left(\frac{1}{\widehat{\sigma}(z)}\right)=O_P(1). \tag{4.14}$$

Combining this results, we conclude that

$$\sup_{\{x,y\}\in D}|M_{\sigma,\widehat{\sigma}}(x,y)|=O_P\left(h_\sigma^{p+1}\right)+O_P\left(\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right).$$

Further, the uniform convergence results for the conventional local polynomial estimator (see Lemma 4.3) provide

$$\frac{1}{n}\sum_{t=1}^{n}\left|(B_{X_{t-k}}+V_{X_{t-k}})(B_{X_{t-l}}+V_{X_{t-l}})\right|\leq\sup_{x\in D}(B_x+V_x)^2=O_P\left(h_0^{p+1}+\frac{\sqrt{\log(n)}}{nh_0}\right)^2,$$

and similarly, using the boundedness of the second moments of $(U_t)_{t\in\mathbb{Z}}$ and Markov's inequality we obtain

$$\frac{1}{n}\sum_{t=1}^{n}\left|(B_{X_{t-k}}+V_{X_{t-k}})U_{t-l}+(B_{X_{t-l}}+V_{X_{t-l}})U_{t-k}\right|=O_P\left(h_0^{p+1}+\frac{\sqrt{\log(n)}}{nh_0}\right).$$

Hence $S_{n,2}=O_P\left(\left(h_\sigma^{p+1}+\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right)\left(h_0^{p+1}+\frac{\sqrt{\log(n)}}{nh_0}\right)\right)$.

Asymptotic order of $S_{n,3}$. Note that

$$\begin{aligned}\frac{\sigma(x)\sigma(y)}{\widehat{\sigma}(x)\widehat{\sigma}(y)}-1&=\left(1+\frac{\sigma(x)-\widehat{\sigma}(x)}{\widehat{\sigma}(x)}\right)\left(1+\frac{\sigma(y)-\widehat{\sigma}(y)}{\widehat{\sigma}(y)}\right)-1\\&=\frac{\sigma(x)-\widehat{\sigma}(x)}{\widehat{\sigma}(x)}+\frac{\sigma(y)-\widehat{\sigma}(y)}{\widehat{\sigma}(y)}+\frac{(\sigma(x)-\widehat{\sigma}(x))(\sigma(y)-\widehat{\sigma}(y))}{\widehat{\sigma}(x)\widehat{\sigma}(y)},\end{aligned}$$

and hence

$$\begin{aligned}
S_{n,3} =& \frac{1}{n}\sum_{t=1}^{n}\left(\frac{\sigma(X_{t-k})\sigma(X_{t-l})}{\widehat{\sigma}(X_{t-k})\widehat{\sigma}(X_{t-l})}-1\right)U_{t-k}U_{t-l} \\
=& \frac{1}{n}\sum_{t=1}^{n}\left(\frac{\sigma(X_{t-k})-\widehat{\sigma}(X_{t-k})}{\widehat{\sigma}(X_{t-k})}+\frac{\sigma(X_{t-l})-\widehat{\sigma}(X_{t-l})}{\widehat{\sigma}(X_{t-l})}\right)U_{t-k}U_{t-l} \\
&+\frac{1}{n}\sum_{t=1}^{n}\left(\frac{(\sigma(X_{t-k})-\widehat{\sigma}(X_{t-k}))(\sigma(X_{t-l})-\widehat{\sigma}(X_{t-l}))}{\widehat{\sigma}(X_{t-k})\widehat{\sigma}(X_{t-l})}\right)U_{t-k}U_{t-l}.
\end{aligned}$$

For the second term, the uniform convergence results for the variance estimator together with Markov's inequality imply

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{t=1}^{n}\left(\frac{(\sigma(X_{t-k})-\widehat{\sigma}(X_{t-k}))(\sigma(X_{t-l})-\widehat{\sigma}(X_{t-l}))}{\widehat{\sigma}(X_{t-k})\widehat{\sigma}(X_{t-l})}\right)U_{t-k}U_{t-l}\right| \\
&\leq \left(\sup_{x\in D}(\sigma(x)-\widehat{\sigma}(x))\right)^2\left(\sup_{x\in D}\left(\frac{1}{\widehat{\sigma}(x)}\right)\right)^2\frac{1}{n}\sum_{t=1}^{n}|U_{t-k}U_{t-l}| \\
&= O_P\left(h_\sigma^{p+1}+\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right)^2.
\end{aligned}$$

The first term requires a more detailed analysis. Decomposition of $1/\widehat{\sigma}(\cdot)$ and repeated usage of the uniform convergence results for the variance estimator $\widehat{\sigma}(\cdot)$ and Markov's inequality provide

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\frac{\sigma(X_{t-k})-\widehat{\sigma}(X_{t-k})}{\widehat{\sigma}(X_{t-k})}U_{t-k}U_{t-l} \\
&= \frac{1}{n}\sum_{t=1}^{n}\frac{\sigma(X_{t-k})-\widehat{\sigma}(X_{t-k})}{\sigma(X_{t-k})}U_{t-k}U_{t-l} \quad +\frac{1}{n}\sum_{t=1}^{n}\frac{(\sigma(X_{t-k})-\widehat{\sigma}(X_{t-k}))^2}{\sigma(X_{t-k})\widehat{\sigma}(X_{t-k})}U_{t-k}U_{t-l} \\
&= \frac{1}{n}\sum_{t=1}^{n}\frac{\sigma(X_{t-k})-\widehat{\sigma}(X_{t-k})}{\sigma(X_{t-k})}U_{t-k}U_{t-l} \quad +O_P\left(h_\sigma^{p+1}+\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right)^2.
\end{aligned}$$

The remaining sum term can be decomposed as follows:

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\frac{\sigma(X_{t-k})-\widehat{\sigma}(X_{t-k})}{\sigma(X_{t-k})}U_{t-k}U_{t-l} \\
&= \frac{1}{n}\sum_{t=1}^{n}\frac{B^\sigma_{X_{t-k}}}{\sigma(X_{t-k})}U_{t-k}U_{t-l} \quad +\frac{1}{n}\sum_{t=1}^{n}\frac{V^\sigma_{X_{t-k}}}{\sigma(X_{t-k})}U_{t-k}U_{t-l}.
\end{aligned}$$

The uniform convergence results for the bias term B^σ_x give

$$\frac{1}{n}\sum_{t=1}^{n}\frac{B^\sigma_{X_{t-k}}}{\sigma(X_{t-k})}U_{t-k}U_{t-l} = O_P\left(h_\sigma^{p+1}\right).$$

We show that for the remaining variance-type term we have

$$\frac{1}{n}\sum_{t=1}^{n}\frac{V^{\sigma}_{X_{t-k}}}{\sigma(X_{t-k})}U_{t-k}U_{t-l} = O_P\left(h_0^{p+1}+\frac{\log(n)}{nh_0}+\frac{1}{\sqrt{n}}\right).$$

The uniform convergence results for $(\mathbf{X}_\sigma^\top \mathbf{W}_\sigma \mathbf{X}_\sigma)^{-1}$ imply that

$$\sup_{x\in D}\left|V_x^\sigma - \mathbf{e}_1^\top \mathbf{M}_\sigma^{-1}/f_X(x)\mathbf{X}_\sigma^\top \mathbf{W}_\sigma \mathbf{V}\right| = o_P(1).$$

Since $\mathbf{M}_\sigma^{-1}$ a finite $(p+1)\times(p+1)-$matrix, and $f_X(x)$ is bounded away from zero, it thus suffices to study the order of

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\left[\mathbf{X}_\sigma^\top \mathbf{W}_\sigma \mathbf{V}\right]_r U_{t-k}U_{t-l}\\
&=\frac{1}{n}\sum_{t=1}^{n}\frac{1}{nh_\sigma}\sum_{i=1}^{n} C_{K_\sigma}(X_i,X_{t-k})\left((m(X_i)-\widehat{m}(X_i))^2+2\sigma(X_i)U_i(m(X_i)-\widehat{m}(X_i))\right.\\
&\quad\left.+\sigma^2(X_i)(U_i^2-1)\right)U_{t-k}U_{t-l},
\end{aligned}$$

where $C_{K_\sigma}(y,z) := K_\sigma\left(\frac{y-z}{h_\sigma}\right)\left(\frac{y-z}{h_\sigma}\right)^{r-1}$, and $r\in\{1,\dots,p+1\}$. Denote

$$\begin{aligned}
V_{n,1} &:= \frac{1}{n}\sum_{t=1}^{n}\frac{1}{nh_\sigma}\sum_{i=1}^{n} C_{K_\sigma}(X_i,X_{t-k})(m(X_i)-\widehat{m}(X_i))^2 U_{t-k}U_{t-l},\\
V_{n,2} &:= \frac{1}{n}\sum_{t=1}^{n}\frac{1}{nh_\sigma}\sum_{i=1}^{n} C_{K_\sigma}(X_i,X_{t-k})U_i(m(X_i)-\widehat{m}(X_i))U_{t-k}U_{t-l},\\
V_{n,3} &:= \frac{1}{n}\sum_{t=1}^{n}\frac{1}{nh_\sigma}\sum_{i=1}^{n} C_{K_\sigma}(X_i,X_{t-k})(U_i^2-1)U_{t-k}U_{t-l}.
\end{aligned}$$

In Lemmas 4.18 – 4.20 we show that $V_{n,i} = O_P\left(h_0^{p+1}+\frac{\log(n)}{nh_0}+\frac{1}{\sqrt{n}}\right),\ i=1,2,3$. Since $\sigma(\cdot)$ is bounded and will hence not effect the asymptotic order, this yields the assertion. □

Lemma 4.18. *Under Assumptions 1 and 3, it holds*

$$V_{n,1} = O_P\left(h_0^{p+1}+\frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right)^2.$$

Proof. The uniform convergence results for the conventional local polynomial estimator given in

Lemma 4.3 together with Markov's inequality provide

$$\begin{aligned}|V_{n,1}| &\leq \sup_{x\in D}(m(x)-\widehat{m}(x))^2\frac{1}{n}\sum_{t=1}^{n}\frac{1}{nh_\sigma}\sum_{i=1}^{n}|C_{K_\sigma}(X_i,X_{t-k})U_{t-k}U_{t-l}|\\ &= O_P\left(h_0^{p+1}+\frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right)^2.\end{aligned}$$

□

Lemma 4.19. *Under Assumptions 1 and 3, it holds*

$$V_{n,2}=O_P\left(h_0^{p+1}+\frac{1}{nh_0}\right).$$

Proof. We have

$$V_{n,2}=\frac{1}{n}\sum_{t=1}^{n}\frac{1}{nh_\sigma}\sum_{i=1}^{n}C_{K_\sigma}(X_i,X_{t-k})U_i(B_{X_i}+V_{X_i})U_{t-k}U_{t-l},$$

where - similar to previous calculations - the uniform convergence results of the bias term combined with Markov's inequality imply

$$\begin{aligned}&\left|\frac{1}{n}\sum_{t=1}^{n}\frac{1}{nh_\sigma}\sum_{i=1}^{n}C_{K_\sigma}(X_i,X_{t-k})U_iB_{X_i}U_{t-k}U_{t-l}\right|\\ &\leq \sup_{x\in D}|B_x|\frac{1}{n}\sum_{t=1}^{n}\frac{1}{nh_\sigma}\sum_{i=1}^{n}|C_{K_\sigma}(X_i,X_{t-k})U_iU_{t-k}U_{t-l}|\\ &= O_P\left(h_0^{p+1}\right).\end{aligned}$$

In what follows, we show that for the variance-type term it holds

$$\frac{1}{n}\sum_{t=1}^{n}\frac{1}{nh_\sigma}\sum_{i=1}^{n}C_{K_\sigma}(X_i,X_{t-k})U_iV_{X_i}U_{t-k}U_{t-l}=O_P\left(\frac{1}{nh_0}\right).$$

With similar convergence arguments for $\left(\mathbf{X}_0^\intercal\mathbf{W}_0\mathbf{X}_0\right)^{-1}$ as for $\left(\mathbf{X}_\sigma^\intercal\mathbf{W}_\sigma\mathbf{X}_\sigma\right)^{-1}$ in the calculations for $S_{n,3}$ in Lemma 4.17 it suffices to study the order of

$$\begin{aligned}T_n :=&\frac{1}{n}\sum_{t=1}^{n}\frac{1}{nh_\sigma}\sum_{i=1}^{n}C_{K_\sigma}(X_i,X_{t-k})U_i\left[\mathbf{X}_0^\intercal\mathbf{W}_0\mathbf{U}\right]_s U_{t-k}U_{t-l}\\ =&\frac{1}{n}\sum_{t=1}^{n}\frac{1}{nh_\sigma}\sum_{i=1}^{n}C_{K_\sigma}(X_i,X_{t-k})U_i\frac{1}{nh_0}\sum_{j=1}^{n}C_{K_0}(X_j,X_i)U_jU_{t-k}U_{t-l},\end{aligned}$$

with $C_{K_0}(y,z):=K_0\left(\frac{y-z}{h_0}\right)\left(\frac{y-z}{h_0}\right)^{s-1}$, for some $s\in\{1,\ldots,p+1\}$ arbitrary.

To make use of the averaging effect, we first study the following centered term

$$\widetilde{T_n} := \frac{1}{n}\sum_{t=1}^{n}\frac{1}{nh_\sigma}\sum_{i=1}^{n}C_{K_\sigma}(X_i,X_{t-k})U_i\frac{1}{nh_0}\sum_{j=1}^{n}C_{K_0}(X_j,X_i)U_j\widetilde{U_t},$$

with $\widetilde{U_t} := U_{t-k}U_{t-l} - \mathbb{E}(U_{t-k}U_{t-l})$. We show that the centered term is of asymptotically negligible order $o_P\left(\frac{1}{nh_0}\right)$. We then show that the difference between this centered and the original term has an order of $O_P\left(\frac{1}{nh_0}\right)$, which yields the assertion.

We have for the centered term

$$\begin{aligned}\mathbb{E}\left(\widetilde{T_n}^2\right) =& \frac{1}{n^6}\sum_{t,s}\sum_{i,j,m,o}\mathbb{E}(U_iU_jU_mU_o\widetilde{U_t}\widetilde{U_s})\\ &\times \mathbb{E}\left(h_\sigma^{-2}h_0^{-2}C_{K_\sigma}(X_i,X_{t-k})C_{K_0}(X_j,X_i)C_{K_\sigma}(X_m,X_{s-k})C_{K_0}(X_o,X_m)\right).\end{aligned}$$

Hoelder's inequality gives

$$\begin{aligned}&\mathbb{E}\left(C_{K_\sigma}(X_i,X_{t-k})C_{K_0}(X_j,X_i)C_{K_\sigma}(X_m,X_{s-k})C_{K_0}(X_o,X_m)\right)\\ &\leq \|C_{K_\sigma}(X_i,X_{t-k})\|_4\|C_{K_\sigma}(X_m,X_{s-k})\|_4\|C_{K_0}(X_j,X_i)\|_4\|C_{K_0}(X_o,X_m)\|_4\\ &= O\left(\sqrt{h_\sigma}\sqrt{h_0}\right) = O(h_0),\end{aligned}$$

since $h_0 \asymp h_\sigma$ by Assumption 3(vi).

The definition of a strongly mixing process (see Definition 2.1) provides that the random process $\left(U_z,U_z,U_z,U_z,\widetilde{U_z},\widetilde{U_z}\right)^{\intercal}_{z\geq 1}$ is strongly mixing with exponentially decreasing α-mixing coefficients

$$\widetilde{\alpha_U}(z) := \begin{cases}\alpha_U(z-|k-l|) & \text{if } z>|k-l|,\\ \alpha_U(0) & \text{otherwise.}\end{cases} \tag{4.15}$$

Further note that $\mathbb{E}(U_z)=0$ and $\mathbb{E}(\widetilde{U_z})=0$ for all $z\in\mathbb{Z}$. Since the ninth moments of $(U_z)_{z\geq 1}$ are finite, we have further

$$\|U_i\|_9\,\|U_j\|_9\,\|U_m\|_9\,\|U_o\|_9\,\|\widetilde{U_t}\|_{9/2}\|\widetilde{U_s}\|_{9/2}\leq C_U<\infty,$$

for $(i,j,m,o,t,s)\in\mathbb{Z}^6$ arbitrary.

Thus setting $k_1=k_2=k_3=k_4=9$, $k_5=k_6=9/2$, and $C_\alpha=1/9$, Corollary 4.15 provides

$$\sum_{i,j,m,o,t-k,s-k}\left|\mathbb{E}\left(U_iU_jU_mU_o\widetilde{U_t}\widetilde{U_s}\right)\right| = O\left(n^3\right).$$

Hence

$$\mathbb{E}\left(\widetilde{T_n}^2\right) = \frac{1}{n^6}\, O\left(h_0^{-3}\right) \sum_{i,j,m,o,t-k,s-k} \left|\mathbb{E}\left(U_iU_jU_mU_o\widetilde{U_t}\widetilde{U_s}\right)\right| = O\left(\frac{1}{n^3h_0^3}\right). \tag{4.16}$$

We are left to study the order of the difference between the centered and the original sum term. It holds for its second moment

$$\begin{aligned}
&\mathbb{E}\left(T_n - \widetilde{T_n}\right)^2 \\
&= \frac{1}{n^6}\sum_{t,s}\sum_{i,j,m,o} \mathbb{E}(U_iU_jU_mU_o)\mathbb{E}(U_{t-k}U_{t-l})\mathbb{E}(U_{s-k}U_{s-l}) \\
&\quad \times \mathbb{E}\left(h_\sigma^{-2}h_0^{-2}C_{K_\sigma}(X_i,X_{t-k})C_{K_0}(X_j,X_i)C_{K_\sigma}(X_m,X_{s-k})C_{K_0}(X_o,X_m)\right).
\end{aligned}$$

This time the averaging effect over t and s does not apply, leading to $O(n)$ more effective summands. However, note that we used a rather rough derivation for the kernel product

$$C_{K_\sigma}(X_i,X_{t-k})C_{K_0}(X_j,X_i)C_{K_\sigma}(X_m,X_{s-k})C_{K_0}(X_o,X_m).$$

A more detailed study will help us to derive a suitable small asymptotic rate. For this purpose, we separate the set $\{1,\ldots,n\}^6$ into two sets of indices I_0 and I_1. I_0 contains all the "harmless" cases where the mean of the kernel product is of order $O\left(h_0^2\right)$. In I_1, we merge all the "critical" cases where the mean has the bigger order $O\left(h_0\right)$.

More precisely, let $I_1 := I_{11} \cup I_{12}$, with

$$I_{11} := \{(i,j,m,o,t-k,s-k) : j = m = t-k, \text{ and } i = o = s-k\}$$

and

$$I_{12} := \{(i,j,m,o,t-k,s-k) : j = o = t-k = s-k, \text{ and } i = m\},$$

and the set of uncritical cases

$$I_0 := \{1,\ldots,n\}^6 \setminus I_1.$$

We start with the set I_1 of critical indices. For $(i,j,m,o,t-k,s-k) \in I_{11}$, Hölder's inequality gives

$$\begin{aligned}
&\mathbb{E}\left(K_\sigma^2\left(\frac{X_i-X_j}{h_\sigma}\right)K_0^2\left(\frac{X_j-X_i}{h_0}\right)\right) \\
&= \mathbb{E}\left(K_\sigma\left(\frac{X_i-X_j}{h_\sigma}\right)K_\sigma\left(\frac{X_j-X_i}{h_\sigma}\right)K_0\left(\frac{X_j-X_i}{h_0}\right)K_0\left(\frac{X_i-X_j}{h_0}\right)\right) \\
&\le O\left(h_\sigma^{1/4}h_\sigma^{1/4}h_0^{1/4}h_0^{1/4}\right) = O(h_0),
\end{aligned}$$

by the assumptions on the bandwidths. Similarly, for $(i, j, m, o, t-k, s-k) \in I_{12}$ we get

$$\mathbb{E}\left(C_{K_\sigma}(X_i, X_j) C_{K_0}(X_j, X_i) C_{K_\sigma}(X_i, X_j) C_{K_0}(X_j, X_i)\right) = O(h_0).$$

Now note that since $(U_t)_{t\in\mathbb{Z}}$ has finite fourth moments, we have

$$\mathbb{E}(U_i U_j U_m U_o) \leq \|U_i\|_4 \|U_j\|_4 \|U_m\|_4 \|U_o\|_4 \leq C,$$

and similarly $\mathbb{E}(U_{t-k}U_{t-l})\mathbb{E}(U_{s-k}U_{s-l}) \leq C$. Since $|I_1| = O(n^2)$, we thus obtain

$$\begin{aligned}
&\frac{1}{n^6} \sum_{(i,j,m,o,t-k,s-k)\in I_1} \mathbb{E}(U_i U_j U_m U_o)\mathbb{E}(U_{t-k}U_{t-l})\mathbb{E}(U_{s-k}U_{s-l}) \\
&\qquad \times \mathbb{E}(h_\sigma^{-2} h_0^{-2} C_{K_\sigma}(X_i, X_{t-k}) C_{K_0}(X_j, X_i) C_{K_\sigma}(X_m, X_{s-k}) C_{K_0}(X_o, X_m)) \\
&= \frac{1}{n^6} O(n^2) O(1)\, O\left(\frac{1}{h_0^3}\right) = O\left(\frac{1}{n^4 h_0^3}\right).
\end{aligned} \tag{4.17}$$

Now let $(i, j, m, o, t-k, s-k) \in I_0$. We first show that for this indices, the expectation of the kernel product possesses the sufficiently small order $O(h_0^2)$. Without loss of generality let $i \neq j$ (if $i = j$, the corresponding mean of the kernel product is zero). We distinguish the cases $j = m$ and $j \neq m$.

If $j = m$, then since $(i, j, m, o, t-k, s-k) \notin I_{11}$, at least one of the following statements holds true:

Case $j \neq t-k$: Since the product densities of $(X_t)_{t\in\mathbb{Z}}$ are bounded by Assumption 1 (iii), and the kernel functions are bounded with compact support by Assumption 3 (iii), we have

$$\begin{aligned}
&\mathbb{E}(C_{K_\sigma}(X_i, X_{t-k}) C_{K_0}(X_j, X_i) C_{K_\sigma}(X_j, X_{s-k}) C_{K_0}(X_o, X_j)) \\
&\leq C\, \mathbb{E}(C_{K_\sigma}(X_i, X_{t-k}) C_{K_0}(X_j, X_i)),
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}(C_{K_\sigma}(X_i, X_{t-k}) C_{K_0}(X_j, X_i)) \\
&= \int_{u,v,w} C_{K_\sigma}(u, v) C_{K_0}(w, u) f_{X_i, X_{t-k}, X_j}(u, v, w)\, du\, dv\, dw \\
&= \int_u \left(\int_{y,z} K_\sigma(y) y^{r-1} K_0(z) z^{s-1} f_{X_{t-k}, X_j | X_i = u}(u - x h_\sigma, u + y h_0) h_\sigma h_0 dy dz \right) f_{X_i}(u)\, du \\
&= O(h_0 h_\sigma) \\
&= O(h_0^2),
\end{aligned}$$

where the last equality follows from Assumption 3 (vi) on the bandwidths.

Case $i \neq o$: Similar calculations as for the previous case lead us to

$$\begin{aligned}&\mathbb{E}(C_{K_\sigma}(X_i, X_{t-k})C_{K_0}(X_j, X_i)C_{K_\sigma}(X_j, X_{s-k})C_{K_0}(X_o, X_j))\\&\leq C\,\mathbb{E}(C_{K_0}(X_j, X_i)C_{K_0}(X_o, X_j)) = O(h_0^2).\end{aligned}$$

Case $i \neq s-k$: We have

$$\begin{aligned}&\mathbb{E}(C_{K_\sigma}(X_i, X_{t-k})C_{K_0}(X_j, X_i)C_{K_\sigma}(X_j, X_{s-k})C_{K_0}(X_o, X_j))\\&\leq C\,\mathbb{E}(C_{K_0}(X_j, X_i)C_{K_\sigma}(X_j, X_{s-k})) = O(h_0^2).\end{aligned}$$

Case $o \neq s-k$: We have

$$\begin{aligned}&\mathbb{E}(C_{K_\sigma}(X_i, X_{t-k})C_{K_0}(X_j, X_i)C_{K_\sigma}(X_j, X_{s-k})C_{K_0}(X_o, X_j))\\&\leq C\,\mathbb{E}(C_{K_\sigma}(X_j, X_{s-k})C_{K_0}(X_o, X_j)) = O(h_0^2).\end{aligned}$$

If $j \neq m$, then since $(i, j, m, o, t-k, s-k) \notin I_{12}$, at least one of the following statements holds true: $j \neq o$, $j \neq t-k$, $j \neq s-k$, $o \neq t-k$, $o \neq s-k$, $t \neq s$, or $i \neq m$. In complete analogy to the case $j = m$ we obtain that $\mathbb{E}(C_{K_\sigma}(X_i, X_{t-k})C_{K_0}(X_j, X_i)C_{K_\sigma}(X_j, X_{s-k})C_{K_0}(X_o, X_j)) = O(h_0^2)$ for all these subcases.

The calculations in the proof of Lemma 3.9 provide $\sum_{i,j,m,o} \mathbb{E}(U_iU_jU_mU_o) = O(n^2)$, and thus we get for the order of I_0

$$\begin{aligned}&\frac{1}{n^6}\sum_{(i,j,m,o,t-k,s-k)\in I_0} \mathbb{E}(U_iU_jU_mU_o)\mathbb{E}(U_{t-k}U_{t-l})\mathbb{E}(U_{s-k}U_{s-l})\\&\quad\times \mathbb{E}(h_\sigma^{-2}h_0^{-2}C_{K_\sigma}(X_i, X_{t-k})C_{K_0}(X_j, X_i)C_{K_\sigma}(X_m, X_{s-k})C_{K_0}(X_o, X_m))\\&= \frac{1}{n^6}\,O\left(n^2 \times n^2 \times \frac{1}{h_0^2}\right) = O\left(\frac{1}{n^2h_0^2}\right).\end{aligned} \tag{4.18}$$

Combining equations (4.17) and (4.18), we get $\mathbb{E}\left(\widetilde{T_n} - T_n\right)^2 = O\left(\frac{1}{n^2h_0^2}\right)$, which together with equation (4.16) and Markov's inequality implies $T_n = O_P\left(\frac{1}{nh_0}\right)$. This yields the assertion. □

Lemma 4.20. *Under Assumptions 1 and 3, it holds*

$$V_{n,3} = O_P\left(\frac{1}{\sqrt{n}}\right).$$

Proof. To make use of the averaging effect over U_t, we decompose $V_{n,3}$ as follows:

$$\begin{aligned}
V_{n,3} &= \frac{1}{n}\sum_{t=1}^{n}\frac{1}{nh_\sigma}\sum_{i=1}^{n}C_{K_\sigma}(X_i,X_{t-k})(U_i^2-1)U_{t-k}U_{t-l}\\
&= \frac{1}{n}\sum_{t=1}^{n}\frac{1}{nh_\sigma}\sum_{i=1}^{n}C_{K_\sigma}(X_i,X_{t-k})(U_i^2-1)\widetilde{U_t}\\
&\quad + \frac{1}{n}\sum_{t=1}^{n}\frac{1}{nh_\sigma}\sum_{i=1}^{n}C_{K_\sigma}(X_i,X_{t-k})(U_i^2-1)\mathbb{E}(U_{t-k}U_{t-l})\\
&=: V_{n,31}+V_{n,32},
\end{aligned}$$

with $\widetilde{U_t}$ as for $V_{n,2}$. As for $V_{n,2}$, we prove that the centered term $V_{n,31}$ is of asymptotically negligible order $o_P\left(\frac{1}{\sqrt{n}}\right)$, and the rest term $V_{n,32}$ has an asymptotic order of $O_P\left(\frac{1}{\sqrt{n}}\right)$. It holds for the second moment of the first sum term

$$\begin{aligned}
&\mathbb{E}\left(V_{n,31}^2\right)\\
&= \frac{1}{n^4}\sum_{i,j,t,s}\mathbb{E}\left(\frac{1}{h_\sigma^2}C_{K_\sigma}(X_i,X_{t-k})C_{K_\sigma}(X_j,X_{s-k})\right)\mathbb{E}\left((U_i^2-1)(U_j^2-1)\widetilde{U_t}\widetilde{U_s}\right).
\end{aligned}$$

Hoelder's inequality provides $\mathbb{E}\left(\frac{1}{h_\sigma^2}C_{K_\sigma}(X_i,X_{t-k})C_{K_\sigma}(X_j,X_{s-k})\right)=O\left(\frac{1}{h_\sigma}\right)$. The definition of alpha-mixing directly implies that the process $(U_z^2-1)_{z\geq 1}$ is alpha-mixing with mixing coefficients $\alpha_U(\cdot)$. Consequently, the centered random process $\left((U_z^2-1),(U_z^2-1),\widetilde{U_z},\widetilde{U_z}\right)^{\mathsf{T}}_{z\geq 1}$ is alpha-mixing with the same exponentially decreasing mixing coefficients $\widetilde{\alpha_U}(\cdot)$ as defined in equation (4.15). Since the ninth moments of $(U_t)_{t\in\mathbb{Z}}$ exist, we have for $(i,j,t,s)\in\mathbb{Z}^4$ arbitrary that $\|U_i^2-1\|_{2/9}\|U_j^2-1\|_{2/9}\|\widetilde{U_t}\|_{2/9}\|\widetilde{U_s}\|_{2/9}\leq C$. Thus setting $C_\alpha=1/9$, Corollary 4.16 provides

$$\sum_{i,j,t,s}\left|\mathbb{E}((U_i^2-1)(U_j^2-1)\widetilde{U_t}\widetilde{U_s})\right| = O(n^2),$$

and hence

$$\mathbb{E}\left(V_{n,31}^2\right) = O\left(\frac{1}{n^2h_\sigma}\right). \tag{4.19}$$

For $V_{n,32}$, we have for its second moment

$$
\begin{aligned}
&\mathbb{E}\left(V_{n,32}^2\right) \\
&= \frac{1}{n^4}\sum_t \mathbb{E}(U_{t-k}U_{t-l})^2 \sum_{i,j}\mathbb{E}\left(\frac{1}{h_\sigma^2}C_{K_\sigma}(X_i,X_{t-k})C_{K_\sigma}(X_j,X_{t-k})\right)\mathbb{E}((U_i^2-1)(U_j^2-1)) \\
&\quad + \frac{1}{n^4}\sum_{t\neq s}\mathbb{E}(U_{t-k}U_{t-l})\mathbb{E}(U_{s-k}U_{s-l})\sum_{i,j}\mathbb{E}\left(\frac{1}{h_\sigma^2}C_{K_\sigma}(X_i,X_{t-k})C_{K_\sigma}(X_j,X_{s-k})\right) \\
&\quad \times \mathbb{E}((U_i^2-1)(U_j^2-1)).
\end{aligned}
$$

Since the process $(U_z^2-1)_{z\geq 1}$ is a centered strong mixing process with exponentially decreasing mixing coefficients $\alpha_U(\cdot)$, we have $\sum_{i,j}\mathbb{E}((U_i^2-1)(U_j^2-1)) = O(n)$ in analogy to Lemma 3.6. Further, Hoelder's inequality provides $\mathbb{E}\left(\frac{1}{h_\sigma^2}C_{K_\sigma}(X_i,X_{t-k})C_{K_\sigma}(X_j,X_{t-k})\right) = O\left(\frac{1}{h_\sigma}\right)$. Now let $s\neq t$. Since all product and conditional densities are bounded and the kernels are bounded with compact support by Assumptions 1 and 3, we have for $i\neq j$

$$
\begin{aligned}
&\mathbb{E}\left(C_{K_\sigma}(X_i,X_{t-k})C_{K_\sigma}(X_j,X_{s-k})\right) \\
&= \int_{u,v,w,x} C_{K_\sigma}(u,v)C_{K_\sigma}(w,x)f_{X_i,X_{t-k},X_j,X_{s-k}}(u,v,w,x)\,du\,dv\,dw\,dx \\
&= \int_{v,x}\left(\int_{y,z} K_\sigma(y)y^{r-1}K_\sigma(z)z^{r-1}f_{((X_i,X_j)|(X_{t-k},X_{s-k})=(v,x))}(v+h_\sigma y, x+h_\sigma z)h_\sigma^2\,dy\,dz\right) \\
&\quad \times f_{X_{t-k},X_{s-k}}(v,x)\,dv\,dx \\
&= O\left(h_\sigma^2\right),
\end{aligned}
$$

and thus $\mathbb{E}\left(\frac{1}{h_\sigma^2}C_{K_\sigma}(X_i,X_{t-k})C_{K_\sigma}(X_j,X_{s-k})\right) = O(1)$. Similarly, for $i=j$ we obtain $\mathbb{E}\left(\frac{1}{h_\sigma^2}C_{K_\sigma}(X_i,X_{t-k})C_{K_\sigma}(X_i,X_{s-k})\right) = O(1)$.

Thus

$$
\begin{aligned}
&\mathbb{E}\left(V_{n,32}^2\right) \\
&= \frac{1}{n^4}\left(\sum_t O(1)\,O\left(\frac{1}{h_\sigma}\right) + \sum_{t\neq s}O(1)\right)\sum_{i,j}\mathbb{E}((U_i^2-1)(U_j^2-1)) \\
&= O\left(\frac{1}{n^2h_\sigma}\right) + O\left(\frac{1}{n}\right) = O\left(\frac{1}{n}\right).
\end{aligned}
\tag{4.20}
$$

Equations (4.19) and (4.20) together with Markov's inequality and Assumption 3 (vi) on the bandwidths yield $V_{n,3} = O_P\left(\frac{1}{\sqrt{n}}\right)$. □

With the help of the convergence results for the estimated filtering parameter $\widehat{\boldsymbol{\alpha}}$, we can now give the asymptotic orders for the rest terms $\mathbf{R}_{n,i}$, $i=1,\ldots,7$. The order of the first term $\mathbf{R}_{n,1}$ is a direct consequence of this last result and previous calculations for the homoscedastic case. The asymptotic study of the second term $\mathbf{R}_{n,2}$ is more complicated and requires a decomposition of

$\widehat{U}_t - U_t$ in analogy to the last proof. The asymptotic rates of the remaining terms follow as direct combinations of the other rates.

Lemma 4.21. *Under Assumptions 1 and 3, it holds*

$$\mathbf{R}_{n,1} = o_P\left(\frac{1}{nh}\right).$$

Proof. It holds

$$\begin{aligned} |[\mathbf{R}_{n,1}]_r| &= \left|\sum_{k=1}^{q}(\widehat{\alpha}_k - \alpha_k)\frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1}\sigma(X_t)U_{t-k}\right| \\ &\leq \sum_{k=1}^{q}|\widehat{\alpha}_k - \alpha_k|\left|\frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1}\sigma(X_t)U_{t-k}\right|. \end{aligned}$$

Lemma 4.17 and Assumption 3 (vi) on the bandwidths provide $\max_k |\widehat{\alpha}_k - \alpha_k| = o_P\left(\frac{1}{\sqrt{nh}}\right)$. Further, we have with the same calculations as in Lemma 3.10

$$\begin{aligned} &\mathbb{E}\left(\left(\frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1}\sigma(X_t)U_{t-k}\right)^2\right) \\ &\leq \frac{1}{nh^2}\,\mathbb{E}\left(K^2\left(\frac{X_1 - x}{h}\right)\left(\frac{X_1 - x}{h}\right)^{2r-2}\sigma^2(X_1)\right)\sigma_U^2 \;+ \frac{2}{nh^2}\sum_{i=2}^{n}|\mathbb{E}(U_{1-k}U_{i-k})| \\ &\times \mathbb{E}\left(K\left(\frac{X_1 - x}{h}\right)\left(\frac{X_1 - x}{h}\right)^{r-1} K\left(\frac{X_i - x}{h}\right)\left(\frac{X_i - x}{h}\right)^{r-1}\sigma(X_1)\sigma(X_i)\right). \end{aligned}$$

Since Assumption 3 provides that the kernels are bounded and have compact support, and further the product density and the variance function are bounded, we have

$$\mathbb{E}\left(K^2\left(\frac{X_1 - x}{h}\right)\left(\frac{X_1 - x}{h}\right)^{2r-2}\sigma^2(X_1)\right) = O(h),$$

and, for $i > 1$,

$$\begin{aligned} &\mathbb{E}\left(K\left(\frac{X_1 - x}{h}\right)\left(\frac{X_1 - x}{h}\right)^{r-1} K\left(\frac{X_i - x}{h}\right)\left(\frac{X_i - x}{h}\right)^{r-1}\sigma(X_1)\sigma(X_i)\right) \\ &= h^2\int K(u)u^{r-1}K(v)v^{r-1}\sigma(x+hu)\sigma(x+hv)f_{X_1,X_i}(x+hu, x+hv)\,du\,dv = O(h^2). \end{aligned}$$

Furthermore, the covariance inequality for α-mixing processes (see Lemma 2.5) provides

$$\sum_{i=2}^{n}|\mathbb{E}(U_{1-k}U_{i-k})| \leq 4\sum_{i=1}^{n-1}\alpha_U(i-1)^{1/2}\|U_{1-k}\|_4\|U_{i-k}\|_4 \leq C\sum_{i=0}^{\infty}\alpha_U(i)^{1/2} \leq C,$$

since $\max_{t\in\mathbb{Z}} \|U_t\|_4 < \infty$ by Assumption 1 (iv), and since the strong mixing coefficients are exponentially decreasing and hence summable by Assumption 3 (iv). Thus

$$\mathbb{E}\left(\left(\frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^{r-1}\sigma(X_t)U_{t-k}\right)^2\right) = O\left(\frac{1}{nh}\right)+O\left(\frac{1}{n}\right).$$

Chebyshev's inequality therefore provides $\frac{1}{nh}\sum\limits_{t=1}^{n} K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^{r-1}\sigma(X_t)U_{t-k} = O_P\left(\frac{1}{\sqrt{nh}}\right)$, and thus $\mathbf{R}_{n,1} = o_P\left(\frac{1}{nh}\right)$. □

Lemma 4.22. *Under Assumptions 1 and 3, it holds*

$$\mathbf{R}_{n,2} = O_P\left(h_0^{p+1}+h_\sigma^{p+1}+\frac{\log(n)}{n\sqrt{h_0h_\sigma}}+\frac{\log(n)}{nh_\sigma}\right)+o_P\left(\frac{1}{\sqrt{nh}}\right).$$

Proof. We have

$$[\mathbf{R}_{n,2}]_r = \sum_{k=1}^{q}\alpha_k\frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^{r-1}\sigma(X_t)(\widehat{U}_{t-k}-U_{t-k}).$$

We decompose the difference between the estimated error and its real value as follows:

$$\begin{aligned}
U_{t-k}-\widehat{U}_{t-k} &= U_{t-k}-\frac{\widehat{\sigma(X_t)U_t}}{\widehat{\sigma}(X_t)}\\
&= U_{t-k}-\frac{Y_{t-k}-m(X_{t-k})}{\widehat{\sigma}(X_{t-k})}+\frac{\widehat{m}(X_{t-k})-m(X_{t-k})}{\widehat{\sigma}(X_{t-k})}\\
&= \left(1-\frac{\sigma(X_{t-k})}{\widehat{\sigma}(X_{t-k})}\right)U_{t-k}+\frac{\widehat{m}(X_{t-k})-m(X_{t-k})}{\widehat{\sigma}(X_{t-k})}.
\end{aligned}$$

Substituting this expression, we obtain

$$\begin{aligned}
&-\frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^{r-1}\sigma(X_t)(\widehat{U}_{t-k}-U_{t-k})\\
&=\frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^{r-1}\sigma(X_t)\frac{\widehat{m}(X_{t-k})-m(X_{t-k})}{\widehat{\sigma}(X_{t-k})}\\
&\quad+\frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^{r-1}\sigma(X_t)\left(\frac{\widehat{\sigma}(X_{t-k})-\sigma(X_{t-k})}{\widehat{\sigma}(X_{t-k})}\right)U_{t-k}\\
&=: [S_{n,1}]_r+[S_{n,2}]_r.
\end{aligned}$$

We show that

$$[S_{n,1}]_r = O_P\left(h_0^{p+1}+h_\sigma^{p+1}+\frac{\log(n)}{n\sqrt{h_0h_\sigma}}\right)+o_P\left(\frac{1}{\sqrt{nh}}\right),$$

where we make use of the convergence rate $O_P\left(h_0^{p+1}\right) + o_P\left(\frac{1}{\sqrt{nh}}\right)$ for the conventional polynomial estimator averaged by the kernel (see Chapter 3), and the uniform convergence rates of the conventional local polynomial estimator, and as well for the variance estimator. For the second sum term, decomposition of $1/\widehat{\sigma}(\cdot)$ yields

$$[S_{n,2}]_r = O_P\left(h_\sigma^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right)^2 + O_P\left(h_\sigma^{p+1}\right) + o_P\left(\frac{1}{\sqrt{nh}}\right),$$

which concludes the proof.

Asymptotic order of $S_{n,1}$. We have for the estimated variance

$$\frac{1}{\widehat{\sigma}(x)} = \frac{1}{\sigma(x)} + \frac{\sigma(x) - \widehat{\sigma}(x)}{\sigma(x)\widehat{\sigma}(x)},$$

which leads us to the following decomposition:

$$\begin{aligned}
&[S_{n,1}]_r \\
&= \frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1}\sigma(X_t) \\
&\quad\times\left(\frac{\widehat{m}(X_{t-k}) - m(X_{t-k})}{\sigma(X_{t-k})} + \frac{(\widehat{m}(X_{t-k}) - m(X_{t-k}))(\sigma(X_{t-k}) - \widehat{\sigma}(X_{t-k}))}{\sigma(X_{t-k})\widehat{\sigma}(X_{t-k})}\right).
\end{aligned}$$

The calculations for the first term are similar to those in Lemma 3.11, keeping in mind that the variance function $\sigma(\cdot)$ is both uniformly bounded and bounded away from zero. Hence

$$\begin{aligned}
&\frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1}\sigma(X_t)\frac{\widehat{m}(X_{t-k}) - m(X_{t-k})}{\sigma(X_{t-k})} \\
&= O_P\left(h_0^{p+1}\right) + o_P\left(\frac{1}{\sqrt{nh}}\right).
\end{aligned}$$

For the remaining term, it holds

$$\begin{aligned}
&\left|\frac{1}{nh}\sum_{t=1}^{n} C_K(X_t)\sigma(X_t)\frac{(\widehat{m}(X_{t-k}) - m(X_{t-k}))(\sigma(X_{t-k}) - \widehat{\sigma}(X_{t-k}))}{\sigma(X_{t-k})\widehat{\sigma}(X_{t-k})}\right| \\
&\le \frac{C_{\sigma,2}}{C_{\sigma,1}}\sup_{x\in D}\left(\frac{1}{\widehat{\sigma}(x)}\right)\sup_{x\in D}|\sigma(x) - \widehat{\sigma}(x)|\sup_{x\in D}|\widehat{m}(x) - m(x)|\frac{1}{nh}\sum_{t=1}^{n}|C_K(X_t)|,
\end{aligned}$$

with $C_K(y) := K\left(\frac{y-x}{h}\right)\left(\frac{y-x}{h}\right)^{r-1}$.

From equation (4.14) we get $\sup_{x\in D}\left(\frac{1}{\widehat{\sigma}(x)}\right) = O_P(1)$. Further, Markov's inequality yields

$\frac{1}{nh}\sum_{t=1}^{n}|C_K(X_t)| = O_P(1)$, and thus

$$\begin{aligned}
&\left|\frac{1}{nh}\sum_{t=1}^{n} C_K(X_t)\sigma(X_t)\frac{(\widehat{m}(X_{t-k})-m(X_{t-k}))(\sigma(X_{t-k})-\widehat{\sigma}(X_{t-k}))}{\sigma(X_{t-k})\widehat{\sigma}(X_{t-k})}\right| \\
&\leq O_P(1)\sup_{x\in D}|\sigma(x)-\widehat{\sigma}(x)|\sup_{x\in D}|\widehat{m}(x)-m(x)| \\
&= O_P\left(\left(h_\sigma^{p+1}+\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right)\times\left(h_0^{p+1}+\frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right)\right),
\end{aligned}$$

where we applied the uniform convergence results for the local polynomial estimator and the variance estimator as stated in Lemma 4.3 and Theorem 4.7, respectively.

Asymptotic order of $S_{n,2}$. Decomposing $1/\widehat{\sigma}(\cdot)$, we obtain

$$\begin{aligned}
&[S_{n,2}]_r \\
&= \frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^{r-1}\sigma(X_t)\left(\frac{\widehat{\sigma}(X_{t-k})-\sigma(X_{t-k})}{\sigma(X_{t-k})}\right)U_{t-k} \\
&\quad -\frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^{r-1}\sigma(X_t)\left(\frac{(\widehat{\sigma}(X_{t-k})-\sigma(X_{t-k}))^2}{\widehat{\sigma}(X_{t-k})\sigma(X_{t-k})}\right)U_{t-k}.
\end{aligned}$$

The uniform convergence results of the variance estimator as stated in Theorem 4.7 give the asymptotic order of the second sum term:

$$\begin{aligned}
&\left|\frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^{r-1}\sigma(X_t)\left(\frac{(\widehat{\sigma}(X_{t-k})-\sigma(X_{t-k}))^2}{\widehat{\sigma}(X_{t-k})\sigma(X_{t-k})}\right)U_{t-k}\right| \\
&\leq \frac{C_{\sigma,2}}{C_{\sigma,1}}\sup_{x\in D}\left(\frac{1}{\widehat{\sigma}(x)}\right)\sup_{x\in D}(\widehat{\sigma}(x)-\sigma(x))^2\,\frac{1}{nh}\sum_{t=1}^{n}\left|K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^{r-1}U_{t-k}\right| \\
&= O_P(1)\,O_P\left(h_\sigma^{p+1}+\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right)^2 O_P(1) \\
&= O_P\left(h_\sigma^{p+1}+\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right)^2.
\end{aligned}$$

For the first sum term, we make use again of the uniform convergence results for the bias term B_x^σ to obtain

$$\begin{aligned}
&\frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^{r-1}\sigma(X_t)\left(\frac{\widehat{\sigma}(X_{t-k})-\sigma(X_{t-k})}{\sigma(X_{t-k})}\right)U_{t-k} \\
&= O_P\left(h_\sigma^{p+1}\right)+\frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^{r-1}\sigma(X_t)\frac{V^\sigma_{X_{t-k}}}{\sigma(X_{t-k})}U_{t-k}.
\end{aligned}$$

The remaining term requires a more detailed analysis. In what follows, we show that this term is of negligible order

$$\frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1} \sigma(X_t)\frac{V^{\sigma}_{X_{t-k}}}{\sigma(X_{t-k})}U_{t-k} = o_P\left(\frac{1}{\sqrt{nh}}\right),$$

which yields the assertion of the Lemma. In total analogy to the proofs of Lemma 4.17 and 4.19, making use of the uniform convergence of $(\mathbf{X}_\sigma^\top \mathbf{W}_\sigma \mathbf{X}_\sigma)^{-1}$ and the fact that $f_X(\cdot)$ is bounded away from zero, it suffices to study the order of

$$\begin{aligned}
&\frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1} \frac{\sigma(X_t)}{\sigma(X_{t-k})}\left[\mathbf{X}_\sigma \mathbf{W}_\sigma \mathbf{V}\right]_s U_{t-k} \\
&= \frac{1}{nh}\sum_{t=1}^{n} C_K(X_t)\frac{\sigma(X_t)}{\sigma(X_{t-k})}\frac{1}{nh_\sigma}\sum_{i=1}^{n} C_{K_\sigma}(X_i, X_{t-k}) \\
&\quad \times (\sigma(X_i)(m(X_i) - \widehat{m}(X_i))^2 + 2\sigma(X_i)U_i(m(X_i) - \widehat{m}(X_i)) + (U_i^2 - 1))U_{t-k},
\end{aligned}$$

with $C_{K_\sigma}(y,z) := K_\sigma\left(\frac{y-z}{h_\sigma}\right)\left(\frac{y-z}{h_\sigma}\right)^{s-1}$ for some $s \in \{1, \ldots, p+1\}$ arbitrary, and $C_K(X_t)$ as before. Keeping in mind that $\sigma(\cdot)$ is both bounded and bounded away from zero, the proof is thus completed by establishing the following asymptotic orders:

$$\begin{aligned}
V_{n,1} :=& \frac{1}{nh}\sum_{t=1}^{n} C_K(X_t)\frac{1}{nh_\sigma}\sum_{i=1}^{n} C_{K_\sigma}(X_i, X_{t-k})(m(X_i) - \widehat{m}(X_i))^2 U_{t-k} \\
=& O_P\left(h_0^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right)^2, \\
V_{n,2} :=& \frac{1}{nh}\sum_{t=1}^{n} C_K(X_t)\frac{1}{nh_\sigma}\sum_{i=1}^{n} C_{K_\sigma}(X_i, X_{t-k})U_i(m(X_i) - \widehat{m}(X_i))U_{t-k} \\
=& O_P\left(h_0^{p+1} + \frac{1}{nh_0}\right), \\
V_{n,3} :=& \frac{1}{nh}\sum_{t=1}^{n} C_K(X_t)\frac{1}{nh_\sigma}\sum_{i=1}^{n} C_{K_\sigma}(X_i, X_{t-k})(U_i^2 - 1)U_{t-k} \\
=& o_P\left(\frac{1}{\sqrt{nh}}\right).
\end{aligned}$$

Asymptotic order of $V_{n,1}$ By virtue of the uniform convergence results for $|m(x) - \widehat{m}(x)|$ as provided in Lemma 4.3, we have

$$
\begin{aligned}
|V_{n,1}| &= \left| \frac{1}{nh} \sum_{t=1}^{n} C_K(X_t) \frac{1}{nh_\sigma} \sum_{i=1}^{n} C_{K_\sigma}(X_i, X_{t-k})(m(X_i) - \widehat{m}(X_i))^2 U_{t-k} \right| \\
&\leq \sup_{x \in D} |m(x) - \widehat{m}(x)|^2 \frac{1}{nh} \frac{1}{nh_\sigma} \sum_{i=1}^{n} \sum_{t=1}^{n} |C_K(X_t) C_{K_\sigma}(X_i, X_{t-k}) U_{t-k}| \\
&= O_P \left(h_0^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_0}} \right)^2 .
\end{aligned}
$$

Asymptotic order of $V_{n,2}$ The uniform convergence results for the bias term B_x as provided in Corollary 2.16 give

$$
V_{n,2} = O_P\left(h_0^{p+1}\right) + \frac{1}{nh} \sum_{t=1}^{n} C_K(X_t) \frac{1}{nh_\sigma} \sum_{i=1}^{n} C_{K_\sigma}(X_i, X_{t-k}) U_i V_{X_i} U_{t-k}.
$$

For the variance-type term, making use again of the uniform convergence results for $(\mathbf{X}_0^\top \mathbf{W}_0 \mathbf{X}_0)^{-1}$, it suffices to study the order of the following term

$$
\widetilde{V_{n,2}} := \frac{1}{nh} \sum_{t=1}^{n} C_K(X_t) \frac{1}{nh_\sigma} \sum_{i=1}^{n} C_{K_\sigma}(X_i, X_{t-k}) \frac{1}{nh_0} \sum_{m=1}^{n} U_i C_{K_0}(X_m, X_i) U_m U_{t-k},
$$

with $C_{K_0}(y,z) := K_0\left(\frac{y-z}{h_0}\right)\left(\frac{y-z}{h_0}\right)^{w-1}$, for some $w \in \{1, \dots, p+1\}$ arbitrary. We show that

$$
\widetilde{V_{n,2}} = O_P\left(\frac{1}{nh_0}\right).
$$

It holds for it second moment:

$$
\begin{aligned}
&\mathbb{E}\left(\widetilde{V_{n,2}}\right)^2 \\
&= \frac{1}{n^6 h^2 h_\sigma^2 h_0^2} \sum_{i,j,m,o,t,s} \mathbb{E}(U_i U_j U_m U_o U_{t-k} U_{s-k}) \times \\
&\quad \mathbb{E}\left(C_K(X_t) C_K(X_s) C_{K_\sigma}(X_i, X_{t-k}) C_{K_\sigma}(X_j, X_{s-k}) C_{K_0}(X_m, X_i) C_{K_0}(X_o, X_j)\right).
\end{aligned} \tag{4.21}
$$

In what follows, we perform a case study concerning the term

$$
\begin{aligned}
&\mathbb{E}\left(C_K(X_t) C_K(X_s) C_{K_\sigma}(X_i, X_{t-k}) C_{K_\sigma}(X_j, X_{s-k}) C_{K_0}(X_m, X_i) C_{K_0}(X_o, X_j)\right) \\
&=: \mathbb{E}(C_{K^2 K_\sigma^2 K_0^2}(i,j,m,o,t-k,s-k)).
\end{aligned}
$$

As in the proof of Lemma 4.17, we separate the set of all indices $\{1, \dots, n\}^6$ into a set of "harmless" indices I_0 and another one of "critical" indices I_1. Here, the set I_0 contains all the "harmless" cases where $\mathbb{E}(C_{K^2K_\sigma^2K_0^2}(i,j,m,o,t-k,s-k)) = O\left(h^2h_0 + hh_0^2\right)$. In I_1, we merge all the "critical" cases where the expectation of the kernel product has the bigger order $O\left(hh_\sigma\right)$. More precisely, let $I_1 := \{(i,j,m,o,t,s) : t-k = s-k = o = m, \text{ and } i = j\}$, and consequently $I_0 = \{1,\ldots,n\}^6 \setminus I_1$.

We start with $(i,j,m,o,t,s) \in I_1$. Note that $|I_1| = O\left(n^2\right)$. Further, it holds for the expectation of the associated kernel product if $i \neq t$[6]

$$\begin{aligned}
&\mathbb{E}(C_K^2(X_t)C_{K_\sigma}^2(X_i,X_{t-k})C_{K_0}^2(X_{t-k},X_i))\\
&\leq C\ \mathbb{E}(C_K^2(X_t)C_{K_\sigma}^2(X_i,X_{t-k}))\\
&= C\ \left(\int_{u,v,w} C_K^2(u)C_{K_\sigma}^2(v,w)f_{X_t,X_i,X_{t-k}}(u,v,w)\ du\ dv\ dw\right)\\
&= C\left(\int_{y,z} K^2(y)y^{2r-2}K_\sigma^2(z)z^{2s-2}f_{X_t,X_i|X_{t-k}=w}(x-hy,w+h_\sigma z)h\ h_\sigma\ dy\ dz\right.\\
&\quad\left.\times\int_w f_{X_{t-k}}(w)dw\right)\\
&= O(hh_\sigma),
\end{aligned}$$

since by Assumptions 1 (iii) and 3 (v) all densities of the design process are bounded, and all kernel functions are bounded with compact support.

From Hoelder's inequality and the existence of the sixth moments of $(U_t)_{t\in\mathbb{Z}}$ by Assumption 3 (iii) we have $|\mathbb{E}(U_iU_jU_mU_oU_{t-k}U_{s-k})| < \infty$, and thus

$$\begin{aligned}
&\frac{1}{n^6h^2h_\sigma^2h_0^2}\sum_{(i,j,m,o,t,s)\in I_1}\mathbb{E}(U_iU_jU_mU_oU_{t-k}U_{s-k})\mathbb{E}\left(C_{K^2K_\sigma^2K_0^2}(i,j,m,o,t-k,s-k)\right)\\
&= \frac{1}{n^6h^2h_\sigma^2h_0^2}\sum_{t,i}O(1)\mathbb{E}(C_K^2(X_t)C_{K_\sigma}^2(X_i,X_{t-k})C_{K_0}^2(X_{t-k},X_i))\\
&= \frac{1}{n^6h^2h_\sigma^2h_0^2}O(n^2)O(1)O(hh_\sigma)\\
&= O\left(\frac{1}{n^4hh_\sigma h_0^2}\right)\\
&= o\left(\frac{1}{n^2h_0^2}\right),
\end{aligned} \tag{4.22}$$

where the last equality follows from Assumption 3 (vi) on the bandwidths.

In what follows, we show briefly that for the remaining indices $(i,j,m,o,t,s) \in I_0$ the associated kernel product has the order $O\left(h^2h_0\right)$. We only give the orders of convergence here, as their calculation is similar to the one above. We make use of the fact, that by Assumption 3 all kernels are bounded with compact support and further $h_0 \asymp h_\sigma$. We distinguish the different cases:

[6]Similarly, for $i = t$ we obtain the order $O(hh_\sigma)$.

- For $s \neq t$, without loss of generalization let $s \neq t-k$.[7] We then have

$$\begin{aligned}
&\mathbb{E}(C_K(X_t)C_K(X_s)C_{K_\sigma}(X_i,X_{t-k})C_{K_\sigma}(X_j,X_{s-k})C_{K_0}(X_m,X_i)C_{K_0}(X_o,X_j))\\
&\leq C\ \mathbb{E}(C_K(X_t)C_K(X_s)C_{K_\sigma}(X_i,X_{t-k}))\\
&= O\left(h^2 h_\sigma\right) = O\left(h^2 h_0\right).
\end{aligned}$$

- For $t-k \neq m$, we have

$$\begin{aligned}
&\mathbb{E}(C_K(X_t)C_K(X_s)C_{K_\sigma}(X_i,X_{t-k})C_{K_\sigma}(X_j,X_{s-k})C_{K_0}(X_m,X_i)C_{K_0}(X_o,X_j))\\
&\leq C\ \mathbb{E}(C_K(X_t)C_{K_\sigma}(X_i,X_{t-k})C_{K_0}(X_m,X_i))\\
&= O\left(h h_\sigma h_0\right) = O\left(h h_0^2\right).
\end{aligned}$$

  The cases $t-k \neq o$, $s-k \neq m$, and $s-k \neq o$ are treated in complete analogy.

- For $m \neq o$, we have either $m \neq t-k$ or $o \neq t-k$. Either way, the expectation of the kernel product has the order $O\left(hh_0^2\right)$ as shown before.

- For $i \neq j$, we have

$$\begin{aligned}
&\mathbb{E}(C_K(X_t)C_K(X_s)C_{K_\sigma}(X_i,X_{t-k})C_{K_\sigma}(X_j,X_{s-k})C_{K_0}(X_m,X_i)C_{K_0}(X_o,X_j))\\
&\leq C\ \mathbb{E}(C_K(X_t)C_{K_\sigma}(X_j,X_{t-k})C_{K_\sigma}(X_i,X_{t-k}))\\
&= O\left(h h_\sigma^2\right) = O\left(h h_0^2\right).
\end{aligned}$$

Note that the process $(U_z, U_z, U_z, U_z, U_{z-k}, U_{z-k})^\intercal_{z\in\mathbb{Z}}$ is centered and α-mixing with exponentially decreasing mixing coefficients $\widetilde{\alpha_U}(\cdot)$ defined as follows:

$$\widetilde{\alpha_U}(z) := \begin{cases} \alpha_U(z-k), & \text{if } z \geq k,\\ \alpha_U(0) & \text{if } z < k. \end{cases} \tag{4.23}$$

Further, the existence of the seventh moments of $(U_t)_{t\in\mathbb{Z}}$ provides

$$\|U_i\|_7\|U_j\|_7\|U_m\|_7\|U_o\|_7\|U_{t-k}\|_7\|U_{s-k}\|_7 \leq C_U < \infty.$$

We can thus apply Lemma 4.14 with $C_\alpha = 1/7$ to obtain

$$\sum_{i,j,m,o,t,s} \mathbb{E}(U_iU_jU_mU_oU_{t-k}U_{s-k}) = O(n^3).$$

[7]For $s = t-k$, since $s \neq t$ we have $t \neq s-k$ and the order of the product term is calculated analogously.

We hence obtain

$$
\begin{aligned}
&\frac{1}{n^6h^2h_\sigma^2h_0^2}\sum_{(i,j,m,o,t,s)\in I_0}\mathbb{E}(U_iU_jU_mU_oU_{t-k}U_{s-k})\times\\
&\quad\mathbb{E}\left(C_K(X_t)C_K(X_s)C_{K_\sigma}(X_i,X_{t-k})C_{K_\sigma}(X_j,X_{s-k})C_{K_0}(X_m,X_i)C_{K_0}(X_o,X_j)\right) \qquad (4.24)\\
&=\frac{1}{n^3}O\left(\frac{1}{h_0h_\sigma^2}+\frac{1}{hh_\sigma^2}\right)=o\left(\frac{1}{n^2h_0^2}\right),
\end{aligned}
$$

again by Assumption 3 (vi) on the bandwidths.

Combining equations (4.21), (4.22) and (4.24) we get $\mathbb{E}\left(\widetilde{V_{n,2}}^2\right)=O\left(\frac{1}{n^2h_0^2}\right)$, and using Markov's inequality we conclude

$$
\widetilde{V_{n,2}}=O_P\left(\frac{1}{nh_0}\right).
$$

Asymptotic order of $V_{n,3}$ We have for the second moment of $V_{n,3}$

$$
\begin{aligned}
\mathbb{E}(V_{n,3}^2)=&\frac{1}{n^4h^2h_\sigma^2}\sum_{i,j=1}^{n}\sum_{t,s=1}^{n}\mathbb{E}\left(C_K(X_t)C_{K_\sigma}(X_i,X_{t-k})C_K(X_s)C_{K_\sigma}(X_j,X_{s-k})\right)\\
&\times\mathbb{E}((U_i^2-1)U_{t-k}(U_j^2-1)U_{s-k}).
\end{aligned}
$$

Note that the process $((U_z^2-1),(U_z^2-1),U_{z-k},U_{z-k})^\intercal_{z\in\mathbb{Z}}$ is centered and strongly mixing with the same exponentially decreasing mixing coefficients $\widetilde{\alpha_U}(\cdot)$ as defined in equation (4.23). Since by Assumption 1 (iii) the seventh moments of the error process $(U_z)_{z\geq 0}$ exist, we have $\|U_i^2-1\|_3\|U_{t-k}\|_7\|U_j^2-1\|_3\|U_{s-k}\|_7\leq C_U<\infty$. We can hence apply Lemma 4.16 with $C_\alpha=1/21$ to get

$$
\sum_{i,j,t,s}\mathbb{E}((U_i^2-1)U_{t-k}(U_j^2-1)U_{s-k})=O\left(n^2\right).
$$

Further, since the kernels $K(\cdot)$ and $K_\sigma(\cdot)$ are both bounded, Hoelder's inequality provides

$$
\begin{aligned}
&\mathbb{E}\left(C_K(X_t)C_{K_\sigma}(X_i,X_{t-k})C_K(X_s)C_{K_\sigma}(X_j,X_{s-k})\right)\\
&\leq\sqrt{\mathbb{E}\left(C_K^2(X_t)C_{K_\sigma}^2(X_i,X_{t-k})\right)\mathbb{E}\left(C_K^2(X_s)C_{K_\sigma}^2(X_j,X_{s-k})\right)}\\
&=O(hh_\sigma).
\end{aligned}
$$

Thus by Assumption 3 (vi) on the bandwidths

$$
\mathbb{E}(V_{n,3}^2)=O\left(\frac{1}{n^4h^2h_\sigma^2}\right)O(n^2)O(hh_\sigma)=O\left(\frac{1}{n^2hh_\sigma}\right),
$$

and from Markov's inequality

$$V_{n,3} = O_P\left(\frac{1}{n\sqrt{hh_\sigma}}\right) = o_P\left(\frac{1}{\sqrt{nh}}\right).$$

□

Lemma 4.23. *Under Assumptions 1 and 3, it holds*

$$\mathbf{R}_{n,3} = O_P\left(h_\sigma^{p+1}\right) + o_P\left(\frac{1}{\sqrt{nh}}\right).$$

Proof. The assertion follows from the proof of Lemma 4.22 (see the calculations for $S_{n,2}$ therein). □

Lemma 4.24. *Under Assumptions 1 and 3, it holds*

$$\mathbf{R}_{n,4} = o_P\left(\frac{1}{\sqrt{nh}}\left(h_\sigma^{p+1} + \frac{1}{\sqrt{nh}}\right)\right).$$

Proof. We calculate

$$|[\mathbf{R}_{n,4}]_r| \leq \sum_{k=1}^{q} |\widehat{\alpha}_k - \alpha_k| \left| \frac{1}{nh} \sum_{t=1}^{n} K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1} (\widehat{\sigma}(X_t) - \sigma(X_t))U_{t-k}\right|.$$

Lemma 4.17 together with Assumption 3 (vi) on the bandwidths provide $|\widehat{\alpha}_k - \alpha_k| = o_P\left(\frac{1}{\sqrt{nh}}\right)$. Further, Lemma 4.23 gives $\frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^{r-1}(\widehat{\sigma}(X_t) - \sigma(X_t))U_{t-k} = O_P(h_\sigma^{p+1}) + o_P\left(\frac{1}{\sqrt{nh}}\right)$, which concludes the proof. □

Lemma 4.25. *Under Assumptions 1 and 3, it holds*

$$\mathbf{R}_{n,5} = o_P\left(\frac{1}{\sqrt{nh}}\left(h_0^{p+1} + h_\sigma^{p+1} + \frac{\log(n)}{n\sqrt{h_0 h_\sigma}} + \frac{\log(n)}{nh_\sigma} + \frac{1}{\sqrt{nh}}\right)\right).$$

Proof. The assertion follows combining Lemma 4.17 and Assumption 3 (vi) providing $|\widehat{\alpha}_k - \alpha_k| = o_P\left(\frac{1}{\sqrt{nh}}\right)$ with Lemma 4.22. □

Lemma 4.26. *Under Assumptions 1 and 3, it holds*

$$\mathbf{R}_{n,6} = O_P\left(h_0^{p+1} + h_\sigma^{p+1} + \frac{\log(n)}{nh_\sigma} + \frac{\log(n)}{n\sqrt{h_0 h_\sigma}}\right).$$

Proof. We have

$$|[\mathbf{R}_{n,6}]_r| \leq \sup_{x\in D} |\widehat{\sigma}(x) - \sigma(x)| \frac{1}{nh}\sum_{t=1}^{n}\left| K\left(\frac{X_t - x}{h}\right)\left(\frac{X_t - x}{h}\right)^{r-1} \sum_{k=1}^{q} \alpha_k(\widehat{U}_{t-k} - U_{t-k})\right|.$$

Theorem 4.7 provides $\sup_{x \in D} |\widehat{\sigma}(x) - \sigma(x)| = O_P\left(h_\sigma^{p+1} + \frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right)$.

We use the same decomposition as in Lemma 4.22 for the remaining term:

$$\begin{aligned}
&\frac{1}{nh}\sum_{t=1}^{n}\left|K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^{r-1}(\widehat{U}_{t-k}-U_{t-k})\right| \\
&\leq \frac{1}{nh}\sum_{t=1}^{n}\left|K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^{r-1}\frac{B_{X_{t-k}}+V_{X_{t-k}}}{\widehat{\sigma}(X_{t-k})}\right| \\
&\quad + \frac{1}{nh}\sum_{t=1}^{n}\left|K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^{r-1}\left(\frac{\widehat{\sigma}(X_{t-k})-\sigma(X_{t-k})}{\widehat{\sigma}(X_{t-k})}\right)U_{t-k}\right| \\
&\leq \sup_{x\in D}\left(\frac{1}{\widehat{\sigma}(x)}\right)\sup_{x\in D}|B_x+V_x|\frac{1}{nh}\sum_{t=1}^{n}\left|K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^{r-1}\right| \\
&\quad + \sup_{x\in D}\left(\frac{1}{\widehat{\sigma}(x)}\right)\sup_{x\in D}|\widehat{\sigma}(x)-\sigma(x)|\frac{1}{nh}\sum_{t=1}^{n}\left|K\left(\frac{X_t-x}{h}\right)\left(\frac{X_t-x}{h}\right)^{r-1}U_{t-k}\right| \\
&= O_P(1)\,O_P\left(h_0^{p+1}+\frac{\sqrt{\log(n)}}{\sqrt{nh_0}}\right)O_P(1) + O_P(1)\,O_P\left(h_\sigma^{p+1}+\frac{\sqrt{\log(n)}}{\sqrt{nh_\sigma}}\right)O_P(1),
\end{aligned}$$

where we applied the same uniform convergence results as in Lemma 4.22 together with Markov's inequality for the sum terms. ☐

Lemma 4.27. *Under Assumptions 1 and 3, it holds*

$$\mathbf{R}_{n,7} = o_P\left(\frac{1}{\sqrt{nh}}\left(h_0^{p+1} + h_\sigma^{p+1} + \frac{\log(n)}{nh_\sigma} + \frac{\log(n)}{n\sqrt{h_0 h_\sigma}}\right)\right).$$

Proof. The assertion follows from Lemmas 4.17 and 4.26 combined with Assumption 3 (vi) on the bandwidths. ☐

Chapter 5

Numerical results

The aim of this chapter is to give empirical evidence that our estimator achieves significant efficiency gains on finite data sets. In the first instance we address the issue of choosing the right model parameters. In the second part of this chapter we investigate our proposed homoscedastic and heteroscedastic estimators on simulated data and compare it with the conventional estimator for a variety of linear and nonlinear error processes.

Contents

5.1 Choice of model

5.1.1 Bandwidth selection

We will first focus on the selection of the main bandwidth h, and then discuss the choice of the prior bandwidths h_0 and h_1. We consider a simple plug-in method, k-fold Cross Validation and Silverman's rule-of-thumb.

The bandwidth controls the complexity of the fit and the right choice of the bandwidth is crucial for a precise estimation. There is a trade-off between variance and bias. Large values for h reduce the variance, since more points are included in the estimate. However, as h increases, the average distance between these "local" points and the point of interest x also increases. This typically results in a larger bias. Minimizing the mean squared error (MSE) is a natural way to deal with this trade-off.

Recall that in order to minimize the MSE, the optimal main bandwidth would be of order $n^{-\frac{1}{2p+3}}$. More precisely, Corollary 4.13 provides that in the heteroscedastic case the leading bias and variance terms are $B(x) = h^{p+1}\frac{m^{(p+1)}(x)}{(p+1)!}\left[\mathbf{M}^{-1}\widetilde{\mathbf{B}}\right]_{1,1}$ and $V(x) = \frac{1}{nh}\sigma^2_{\overline{U}^{\text{het}}}\frac{\sigma^2(x)}{f_X(x)}\left[\mathbf{M}^{-1}\mathbf{\Gamma}\mathbf{M}^{-1}\right]_{1,1}$ respectively. Balancing these two properties, we obtain a local optimal bandwidth

$$h^{\text{opt}}(x) = C_p(K)\left(\frac{\sigma^2_{\overline{U}^{\text{het}}}\sigma^2(x)}{f_X(x)\left(\frac{m^{(p+1)}(x)}{(p+1)!}\right)^2}\right)^{\frac{1}{2p+3}} n^{-\frac{1}{2p+3}}, \tag{5.1}$$

where $C_p(K)$ summarizes all quantities that only depend on the kernel function $K(\cdot)$.

In most applications, one is more interested in a global bandwidth choice. A common measure for a global bandwidth choice is the weighted integrated mean squared error (IMSE), being defined as

$$\text{IMSE} := \int (B^2(x) + V(x))w(x)dx.$$

Here, $w(\cdot)$ denotes a positive weight function. Minimization of the weighted IMSE leads to a global optimal bandwidth choice of

$$h^{\text{globopt}} = C_p(K)\left(\frac{\sigma^2_{\overline{U}^{\text{het}}}\int_D \sigma^2(x)w(x)dx}{\int_D f_X(x)\left(\frac{m^{(p+1)}(x)}{(p+1)!}\right)^2 w(x)dx}\right)^{\frac{1}{2p+3}} n^{-\frac{1}{2p+3}}, \tag{5.2}$$

with D the range in which the estimation is conducted. The respective bandwidth choices in the homoscedastic case are obtained by setting $\sigma^2(x) \equiv 1$.

These bandwidth formulas are identical to that for the conventional local polynomial estimator (see for example Fan & Gijbels (1996, p. 67)) except that the smaller variance $\sigma^2_{\overline{U}^{\text{het}}}$ appears here. Hence, any plug-in method defined for the usual estimators can be applied here with some simple modification in order to estimate the unknown quantities in equation (5.1). For example, a nonparametric method could be applied to estimate $f_X(x)$ and $\sigma^2(x)$.

Alternatively, we could use one of the bandwidth selection methods in the independent setting. Fan & Gijbels (1995) introduced a fully automatic data - driven local bandwidth selection method. Ruppert & Wand (1994) and Ruppert et al. (1995) developed an empirical-bias bandwidths selection algorithm which performs well both asymptotically and in practice. Recently, Zhang et al. (2008) applied an intersection confidence interval method which outperforms other plug-in methods in many signal processing settings.

All these methods are very useful, but their complexity is usually high. Since our proposed estimation method requires us to select three unknown bandwidths, and our main purpose is to compare different local polynomial methods, we will thus stick to an easy plug-in method: A simple rule of thumb (Fan & Gijbels 1996, p. 110), which has proven to perform well if the true curve does not show many high-frequency alternations. Setting $\sigma_{\widetilde{U}} = 1$ (alternatively, one can estimate this value and multiply the bandwidth by it), and taking the weight function $w(x) = f_X(x)1_{[a,b]}(x)$, this method calculates the bandwidth as follows:

1. Fit a polynomial of order $p+3$ globally to the data, leading to the a-priori estimate
$$\breve{m}(x) = \breve{\eta}_0 + \ldots + \breve{\eta}_{p+3}\, x^{p+3}.$$
2. Denote the standardized mean squared error of this estimate by $\breve{\sigma}^2$.
3. An estimator of the $(p+1)$th derivative of $m(\cdot)$ is given by
$$\breve{m}^{(p+1)}(x) = (p+1)!\breve{\eta}_{p+1} + (p+2)!\breve{\eta}_{p+2}\, x + \frac{(p+3)!}{2}\breve{\eta}_{p+3}\, x^2,$$
which is of a quadratic form, allowing for a certain flexibility in estimating the curvature.
4. Regarding the conditional variance $\sigma^2(x)$ as a constant σ^2 with its estimate $\breve{\sigma}^2$, and substituting the estimates into the asymptotically optimal constant bandwidth formula, we obtain the rule of thumb bandwidth selector
$$h_{\text{plug in}} = C_p(K) \left(\frac{\breve{\sigma}^2(b-a)}{\sum_{i=1}^n \left(\breve{m}^{(p+1)}(X_i)\right)^2 1_{[a,b]}(X_i)} \right)^{\frac{1}{(2p+3)}},$$
where we approximated the denominator by an averaged sum.

Note that for standard kernel functions, such as the Gaussian or the Epanechnikov kernel, the quantity $C_p(K)$ is comparatively easy to calculate and is listed for example in Gijbels & Fan (1995).

Another common approach to a global bandwidth choice is cross-validation (CV). The presence of large datasets makes k-fold cross-validation attractive for our simulation analysis. In this approach, the dataset $D := (X_t)_{t=1,\ldots,n}$ is randomly partitioned in k mutually exclusive subsets $D_1, \ldots, D_k$ (the *folds*) of approximately equal size. Of the k subsamples, a single subsample is used as validation data, and the remaining $k-1$ subsamples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data. The average squared error of the k repetitions then gives the cross-validation estimate of the mean-squared prediction error.

More precisely, denote by $\widetilde{m}^{(-M_i)}(\cdot)$ our proposed estimator, computed with the ith part of the data removed. Then the cross-validation estimate of the mean-squared prediction error is

$$CV(h) := \frac{1}{n} \sum_{i=1}^{k} \sum_{t \in M_i} \left(Y_t - \widetilde{m}^{(-M_i)}(X_t) \right)^2 .$$

We now select the CV-bandwidth as the value that minimizes the cross-validation error

$$h_{CV} := \arg \min_{h \in H} CV(h),$$

where H is some set of possible values for h. Regarding the datasets in our simulation study, we choose $k = 10$ and H a grid of the interval $I = \left[0.1 s_X \, n^{-\frac{1}{2p+3}}, 5 s_X \, n^{-\frac{1}{2p+3}} \right]$, where s_X denotes the sample standard deviation. For sake of simplicity, a weighted cross-validation was not taken into account.

As a reference bandwidth choice, we use Silverman's Rule-of-Thumb (short: Silverman's ROT, Silverman (1986)) for density estimation: the optimal bandwidth when estimating a one - dimensional density is $1.06 s_X \, n^{-\frac{1}{2p+3}}$ for the Gaussian kernel, and $2.34 s_X \, n^{-\frac{1}{2p+3}}$ for the Epanechnikov kernel, respectively. This approach is a frequently used bandwidth choice in many univariate regression problems.

Table 5.1 reports the empirical IMSE for $m(x) = x$ with $X_t \sim \mathcal{U}[-2, 2]$ i.i.d, a Gaussian Kernel, and a smoothing degree of $p = 3$ for different sizes n of the data set and a simple AR(1) error process $U_t = 0.3 \, U_{t-1} + \varepsilon_t$, with $\varepsilon_t \sim \mathcal{N}(0, 0.25)$ i.i.d. We choose the prior bandwidths $h_0 = h_\sigma = h^{1.1}$. Here, the empirical IMSE denotes the mean of average squared errors over all data points $X_1, \ldots, X_n$, and a number of 100 replications.[1]

	Empirical IMSE		
	Silverman's ROT	Plug-in	CV
$n = 100$	0.0170	0.0244	0.0205
$n = 200$	0.0139	0.0140	0.0109
$n = 500$	0.0027	0.0048	0.0016
$n = 1000$	0.0014	0.0020	0.0009

Table 5.1: Bandwidth selection: Empirical IMSE for different bandwidth choices and data sizes.

It results that Silverman's rule-of-thumb rules out the other two bandwidth choices for small datasets ($n = 100$), whereas for a large number of observations one should stick to CV. We also implemented CV with more folds ($k = n$) for $n = 100$ data points, but this did not cause a noticeable smaller IMSE. The plug-in method used here seems to be a somewhat to rough approach to match the true optimal bandwidth.

To obtain our proposed estimator, we also need to choose the prior bandwidths h_0 and h_σ. As demonstrated in Chapter 4, if we were given a main bandwidth choice $\widehat{h}$ for h, we could choose

[1] The different bandwidth choices were also tested on other error processes and similar results were obtained.

the prior bandwidths to be $h_0 = \widehat{h}^{\delta_0}$ and $h_\sigma = \widehat{h}^{\delta_1}$, for some $1 < \delta_1, \delta_2 < 2.5$ in order to satisfy all bandwidth conditions.

Another approach of choosing the prior bandwidth is to balance the second order terms. In the homoscedastic case, these terms are of orders h_0^{p+1}, $\frac{1}{n\sqrt{hh_0}}$, and $\frac{h^{p+1}}{\sqrt{nh_0}}$. Balancing those terms for $h \asymp n^{-\frac{1}{2p+3}}$ would give an order of $h_0 \asymp n^{-\frac{1}{2p+3}} h^{\frac{2p+2}{2p+3}} \asymp n^{-\frac{1}{2p+3} - \frac{2p+2}{(2p+3)^2}}$. Now, similar plug-in methods as for the main bandwidth could be applied in order to approximate the asymptotic optimal bandwidth constant. Alternatively, one could use Silverman's rule of thumb. Of course, one could also apply CV for choosing the prior bandwidths, though this heavily effects the runtime especially for bigger datasets.

However, our simulations show that the efficiency gain is not sensitive to the exact choice of the prior bandwidths, so that we stick to the simple choice: $h_0 = h_\sigma = h^{1.1}$.

5.1.2 Other model parameters

Since our main focus is to compare the efficiency of our proposed estimator with the conventional one, we will not try to optimize these estimators. We rather stick to common choices for the kernel functions and the order of fit p. More precisely, we use the Epanechnikov kernel $K(x) := |0.75(1-x^2)|1_{|x|\leq 1}$, due to its optimality in terms of asymptotic IMSE (Wand & Jones 1995). Concerning the smoothing degree, we choose $p = 3$, having in mind that odd degrees of smoothing have proven to provide a smaller variability (Fan & Gijbels 1996). Other kernels and smoothing parameters were also tried and qualitatively similar results were obtained.[2]

In order to verify suitable choices for the AR parameter q, we test our proposed estimator on the regression function $m(x) = x$ and the following strongly mixing linear and nonlinear error-processes:[3]

$$\begin{aligned}
\text{AR(2)} : U_t &= 0.3\, U_{t-1} + 0.2\, U_{t-2} + \varepsilon_t, \\
\text{AR(3)} : U_t &= 0.2\, U_{t-1} + 0.3\, U_{t-2} + 0.2\, U_{t-3} + \varepsilon_t, \\
\text{MA(1,2)} : U_t &= 0.2\, U_{t-1} + 0.2\, \varepsilon_{t-1} + 0.3\, \varepsilon_{t-2} + \varepsilon_t, \\
\text{NLEP} : U_t &= 0.3\, U_{t-1} + 0.8\, \exp(-0.05\, U_{t-1}^2) + \varepsilon_t,
\end{aligned}$$

with $\varepsilon_t \sim \mathcal{N}(0,1)$ i.i.d.

We compute $n = 100$ data points and 100 replications. Concerning the size of the data set, we use Silverman's rule-of-thumb as bandwidth choice. In Table 5.2 we report the empirical IMSE for different choices of q.

[2]We also implemented the Gaussian, the Uniform, and the Triangular kernel, as well as the orders $1 \leq p \leq 4$.
[3]In Section 5.2.2 we comment on the strong mixing property of NLEP.

	Empirical IMSE			
	$q = 1$	$q = 2$	$q = 3$	$q = 4$
AR(2)	0.0607	0.0608	0.0605	0.0603
AR(3)	0.0528	0.0527	0.0528	0.0529
MA(1,2)	0.0816	0.0820	0.0819	0.0818
NLEP	0.0377	0.0376	0.0377	0.0380

Table 5.2: Choice of filtering parameter: Empirical IMSE for different choices of q and error processes.

Note that we would expect an optimal value of $q = 2$ and $q = 3$ for AR(2) and AR(3), respectively. However, our simulation results give evidence that the goodness of our estimator is not sensitive to the exact choice of q.

Since our experimental design is very regular, we did not concern robust or regularized modifications of our estimators. However, for data sets that are likely to possess various outliers, one might consider a robust modification of the local polynomial estimator using M-estimators (see, e.g. Cleveland (1979), Welsh (1996), Fan & Jiang (2000)). For sparse or clustered data, we recommend the usage of regularized modifications of the local polynomial estimator, such as local polynomial ridge (Seifert & Gasser 1996, 2000), or local polynomial LASSO estimators (Vidaurre Henche et al. 2012). The corresponding robust and regularized modifications of our proposed more efficient estimation procedure are straightforward. However, it is not clear and has to be verified if our asymptotic results remain valid for this kind of modified estimators.

5.2 Efficiency comparison

In this section we study the efficiency gain of our proposed estimator compared to the conventional one on various ARMA(1,2) and nonlinear strongly mixing error processes in both the homoscedastic and the heteroscedastic setting.

We present here the simulation results for

$$m(x) = 5\,\frac{\exp(x)}{1 + \exp(x)},$$

only mentioning that similar conclusions can be drawn for a wide set of test functions $m(\cdot)$. Further, we choose again $X_t \sim \mathcal{U}[-2, 2]$ i.i.d. Figure 5.1 illustrates an example of the simulated data. Concerning the model parameters, we choose a filtering parameter of $q = 2$. As for the bandwidth choice, we use Silverman's rule-of thumb for $n = 100$ and 10-fold cross-validation for a larger number of observations.

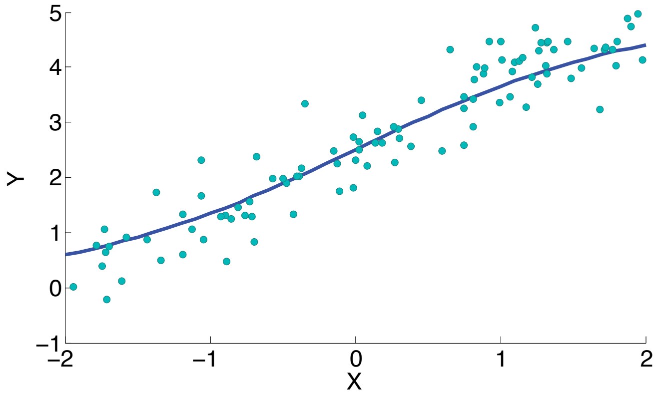

Figure 5.1: Regression function $m(\cdot)$ and generated data: AR(1)-noise with $\alpha = 0.2$ and $n = 100$ observations.

Some general conclusions can be drawn from our simulation experiments:

- The proposed estimator significantly improves the efficiency over the conventional one in most cases apart from the i.i.d. case.
- In general, the more the serial dependence in the error process, the larger the achieved efficiency gain. However, we observe that if the serial dependence is to strong (close to being a degenerate error process), the efficiency gain decreases.
- The empirical relative efficiency improves with growing sample size, especially if the serial dependence is quite strong.
- The highest efficiency gain was observed for error processes of ARMA type with negative coefficients.
- Our results indicate that a significant efficiency gain can be achieved for various types of nonlinear strongly mixing error processes.
- Due to the estimation of the variance function, the proposed heteroscedastic estimator performs quite poor on small datasets ($n = 100$). For a reasonable large number of observations ($n = 1000$) and ARMA error processes with a significant AR(1)-part, it outperforms the conventional as well as the homoscedastic estimator.

The following subsections provide some more detailed results of our experiments. In Tables 5.3–5.7 we report the empirical relative efficiencies (RE) of the proposed estimators over the conventional estimator in terms of IMSE:

$$\text{RE} := \frac{\text{IMSE(homoscedastic estimator)}}{\text{IMSE(conventional estimator)}}, \quad \text{and } \text{RE}^{\text{het}} := \frac{\text{IMSE(heteroscedastic estimator)}}{\text{IMSE(conventional estimator)}}.$$

Here, with IMSEn we denote the empirical IMSE, which is calculated based on the average squared errors over all n data points and a number of 100 replications.

5.2.1 Linear autoregressive error processes

We start our studies with various special cases of the simple AR(1)-process $U_t = \alpha U_{t-1} + \varepsilon_t$, with $\varepsilon_t \sim \mathcal{N}(0, 0.25)$ i.i.d., and $|\alpha| < 1$. Remember from Example 3.2, that the theoretical asymptotic RE in terms of MSE[4] in this case would be $(1-\alpha^2)^{(2p+2)/(2p+3)} = (1-\alpha^2)^{8/9}$. However, our simulations (see Figure 5.2) indicate that if the AR coefficient α is positive, the relative efficiency first improves as it increases and then worsens as it approaches one. This suggests that a different asymptotic theory might apply when α is very close to one, the so-called *unit root case*. Xiao et al. (2003) and Su & Ullah (2006) also commented on this problem. They suggest to choose a larger bandwidth in the region where $\alpha > 0.8$. Surprisingly, this effect does not occur if $\alpha < 0$. In this region, the empirical results for the RE are very close to the theoretical asymptotic RE.

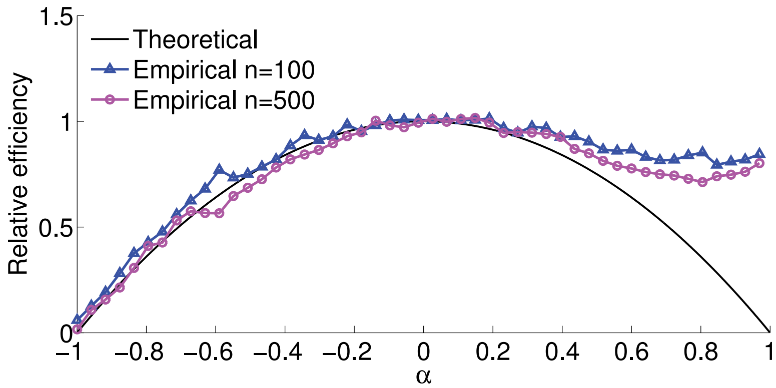

Figure 5.2: AR(1) error processes: Empirical relative efficiency of proposed efficient over conventional estimator for different sample sizes along with its asymptotic value.

We now investigate our estimator on various ARMA(1,2) processes

$$U_t = \alpha U_{t-1} + \varepsilon_t + \gamma_1 \varepsilon_{t-1} + \gamma_2 \varepsilon_{t-2}, \tag{5.3}$$

with $\varepsilon_t \sim \mathcal{N}(0, 0.25)$ i.i.d. In Table 5.3 we report the empirical IMSE and RE for different ARMA-parameters and $n = 100$ and $n = 500$ data points, respectively. The results indicate that our proposed estimator is indeed more efficient than the conventional one in the presence of serial correlation, and the efficiency gain is the largest in the presence of negative serial dependence. In the case of independent errors, our methods is slightly less efficient than the conventional one. This is natural since prewhitening brings no advantage in terms of efficiency in this case but causes computational inaccuracies. The observed RE values also illustrate that the efficiency gain increases with increasing sample size. Note that for error processes with a nonzero MA part ($\gamma_1 \neq 0$ or $\gamma_2 \neq 0$), the suggested filtering procedure is likely to be quite far from matching the true autocorrelation structure. Nevertheless, even in those cases there are positive results.

[4]Note that the asymptotic RE in terms of MSE is the same as measured in IMSE in our simulation experiments.

ARMA parameters			IMSE 100		IMSE 500		Relative Efficiency	
α	γ_1	γ_2	Conv	Eff	Conv	Eff	$n = 100$	$n = 500$
0.0	0.0	0.0	0.0129	0.0132	0.0031	0.0032	1.023	1.032
0.3	0.0	0.0	0.0173	0.0163	0.0040	0.0037	0.942	0.925
-0.3	0.0	-0.2	0.0155	0.0132	0.0036	0.0030	0.852	0.833
0.0	0.0	0.5	0.0210	0.0194	0.0041	0.0036	0.923	0.877
0.0	0.0	-0.5	0.0176	0.0145	0.0038	0.0030	0.822	0.780
0.0	0.9	0.0	0.0242	0.0190	0.0043	0.0032	0.783	0.739
0.0	-0.9	0.0	0.0232	0.0158	0.0055	0.0028	0.679	0.503
0.0	0.7	0.3	0.0253	0.0201	0.0051	0.0038	0.794	0.755
0.8	0.5	0.4	0.3064	0.2337	0.0595	0.0428	0.761	0.719
-0.8	-0.5	-0.4	0.0475	0.0193	0.0101	0.0036	0.407	0.354

Table 5.3: ARMA error processes: Empirical relative efficiencies of proposed efficient (Eff) over conventional estimator (Conv) for different sample sizes.

We also studied ARMA processes of higher orders p and q that are not reported here. The observed efficiency gain was very similar, i.e. the larger the (negative) ARMA parameters and the number of observations, the higher the efficiency gain.

5.2.2 Nonlinear strongly mixing error processes

We now extend our studies to nonlinear strongly mixing autoregressive error processes. We consider the following type of error processes

$$U_t = g(U_{t-1}, \ldots, U_{t-k}) + \varepsilon_t. \tag{5.4}$$

The stationarity and mixing conditions of this type of error processes can be justified using some general results provided by Masry & Tjøstheim (1995, 1997), as mentioned in Example 2.4. Loosely speaking, they found that any nonlinear autoregressive process as specified in equation (5.4) is stationary and strong mixing with exponentially decreasing strong mixing coefficients provided that the growth at infinity of the non-periodic function $g(\cdot)$ is dominated by the growth of a linear function with absolute sum of coefficients smaller than one.

As a first example of this type we consider the following nonlinear error process (NLEP):

$$\text{NLEP}(1): \; U_t = \alpha\, U_{t-2} + \gamma\, U_{t-1} \exp(-0.1\, U_{t-1}^2) + \varepsilon_t,$$

with $\varepsilon_t \sim \mathcal{N}(0, 25)$ i.i.d. In Table 5.4 we report the empirical RE for different values of α and γ and a number of $n = 100$ and $n = 500$ data points, respectively.

Parameters		IMSE 100		IMSE 500		Relative Efficiency	
α	γ	Conv	Eff	Conv	Eff	$n = 100$	$n = 500$
0.0	0.0	0.0129	0.0132	0.0031	0.0032	1.023	1.032
0.0	0.7	0.0212	0.0204	0.0031	0.0028	0.962	0.880
0.0	0.9	0.0261	0.0233	0.0050	0.0043	0.894	0.865
0.0	-0.9	0.0221	0.0211	0.0202	0.0192	0.950	0.942
0.2	0.4	0.0217	0.0202	0.0039	0.0031	0.932	0.801
0.6	0.4	0.0592	0.0471	0.0157	0.0120	0.796	0.766
0.8	0.4	0.2154	0.1819	0.0355	0.0288	0.844	0.812
-0.8	-0.4	0.0169	0.0167	0.0143	0.0140	0.976	0.965

Table 5.4: Nonlinear strongly mixing error processes (1): Empirical relative efficiencies of proposed efficient (Eff) over conventional estimator (Conv) for different sample sizes.

As a second example we consider various cases of the error process

$$\text{NLEP(2)}: \; U_t = \gamma \, U_{t-1}\varphi(U_{t-1}) + \varepsilon_t,$$

with uniform distributed innovations $\varepsilon_t \sim \mathcal{U}[-0.5, 0.5]$ i.i.d. and $\varphi(\cdot)$ the standard normal density function. The obtained results are given in Table 5.5.

Parameters	IMSE 100		IMSE 500		Relative Efficiency	
γ	Conv	Eff	Conv	Eff	$n = 100$	$n = 500$
0.5	0.0056	0.0055	0.0014	0.0013	0.992	0.929
1.0	0.0066	0.0060	0.0016	0.0014	0.911	0.890
2.0	0.0136	0.0110	0.0049	0.0039	0.846	0.805
2.5	0.0310	0.0256	0.0052	0.0043	0.827	0.819
-2.0	0.0229	0.0217	0.0212	0.0207	0.950	0.940

Table 5.5: Nonlinear strongly mixing error processes (2): Empirical relative efficiencies of proposed efficient (Eff) over conventional estimator (Conv) for different sample sizes.

The results give evidence that the proposed estimator outperforms the conventional one even for this type of nonlinear error processes. We notice again that the efficiency gain increases with sample size. Further, we obtain that - up to a certain degree - the more the serial dependence in the error process, the larger is the achieved efficiency gain. However, in contrast to the ARMA error case we obtain that the efficiency gain is actually smaller if the parameters are negative.

5.2.3 Heteroscedastic error processes

We study various special cases of the ARMA(1,2) process as introduced in equation (5.3), now adding a variance function $\sigma(x) = 0.5 \cdot (|x| + 0.5)^2$ to the regression model.[5] Note that the unweighted empirical IMSE is not a reasonable measure of goodness-of-fit in this setting, since

[5]Other variance functions were also implemented and similar results were obtained.

the average MSE is proportional to the variance. For this reason, we use the weighted counterpart of the IMSE with the weighting function $w(x) := 1/\sigma(x)$.

Table 5.6 and 5.7 report the relative efficiencies RE and REHet of our proposed homoscedastic and heteroscedastic estimators - $\widetilde{m}(\cdot)$ and $\widetilde{m}^{\text{het}}(\cdot)$ - over the conventional estimator $\widehat{m}$ respectively for different ARMA parameters $\alpha, \gamma_1, \gamma_2$ and $n \in \{100, 1000\}$ data points, with $\varepsilon_t \sim \mathcal{N}(0, 1)$ i.i.d.

ARMA parameters			Weighted IMSE			Relative Efficiency	
α	γ_1	γ_2	Conv	Eff	EffHet	RE	REHet
0.0	0.0	0.0	0.0492	0.0506	0.0554	1.028	1.126
0.3	0.0	0.0	0.0555	0.0545	0.0601	0.983	1.084
0.9	0.0	0.0	0.5962	0.5066	0.5706	0.850	0.957
-0.6	0.0	0.0	0.0911	0.0742	0.0901	0.814	0.989
-0.9	0.0	0.0	0.2961	0.1732	0.2664	0.586	0.897
0.0	0.9	0.0	0.0945	0.0851	0.0910	0.901	0.963
0.0	0.7	0.3	0.0859	0.0768	0.0878	0.894	1.022
0.8	0.5	0.4	0.9290	0.7861	0.8754	0.846	0.942
-0.8	-0.5	-0.4	0.3292	0.2061	0.3201	0.625	0.975

Table 5.6: Heteroscedastic case (1): Empirical relative efficiency of proposed efficient heteroscedastic (EffHet) and homoscedastic (Eff) over conventional estimator (Conv) in heteroscedastic setting for $n = 100$ data points.

The observed relative efficiencies for $n = 100$ observations indicate that the performance of the proposed heteroscedastic estimator is comparable to that of the conventional one. The proposed homoscedastic estimator outperforms both estimators in the presence of serial dependence. The poor performance of the heteroscedastic estimator for this amount of data is quite natural, since a precise estimation of the variance function requires more observations.

ARMA parameters			Weighted IMSE			Relative Efficiency	
α	γ_1	γ_2	Conv	Eff	EffHet	RE	REHet
0.0	0.0	0.0	0.0009	0.0009	0.0010	1.003	1.023
0.3	0.0	0.0	0.0012	0.0011	0.0013	0.964	1.087
0.9	0.0	0.0	0.0136	0.0114	0.0112	0.841	**0.821**
-0.6	0.0	0.0	0.0016	0.0011	0.0015	0.688	0.981
-0.9	0.0	0.0	0.0043	0.0020	0.0017	0.457	**0.400**
0.0	0.9	0.0	0.0020	0.0018	0.0020	0.882	0.992
0.0	0.7	0.3	0.0018	0.0015	0.0017	0.862	0.944
0.8	0.5	0.4	0.0154	0.0119	0.0106	0.771	**0.688**
-0.8	-0.5	-0.4	0.0049	0.0025	0.0018	0.500	**0.376**

Table 5.7: Heteroscedastic case (2): Empirical relative efficiency of proposed efficient heteroscedastic (EffHet) and homoscedastic (Eff) over conventional estimator (Conv) in heteroscedastic setting for $n = 1000$ data points.

The results for $n = 1000$ data points indicate that in this case the proposed heteroscedastic estimator outperforms the conventional one as well as the homoscedastic estimator for high values of α (for a better reading, we highlighted those RE-values where the heteroscedastic estimator performs best). However, we acknowledge that the performance of the heteroscedastic estimator for moderate sized parameters α is still quite poor.

Chapter 6

Conclusion and Discussion

We adapted a modification of local polynomial time series fitting introduced by Xiao et al. (2003) to the case of strongly mixing observation errors. We demonstrated that this method asymptotically improves the efficiency over the conventional estimator. The improvement depends only on the short-run variance of the error process and can thus be arbitrarily large. However, for error processes with a nonlinear autocorrelation structure, it is a rather complicated task to derive the error variance theoretically. In such cases, one can approximate the efficiency gain empirically by the sample variance.

We further proposed a more efficient three-step estimator for nonparametric regression with observation errors that are strongly mixing and heteroscedastic. In this context, we were able to derive uniform convergence results for the estimator of the conditional variance.

Simulations confirmed the finite sample out-performance of our estimator over the conventional one under serial correlation and heteroscedasticity for error processes with both linear and nonlinear serial correlation. Surprisingly, the goodness of our estimator is not at all sensitive to the choice of the filtering parameter q. One can expect that the finite sample performance might be even further improved by a more sophisticated bandwidth choice (see Chapter 5.1 for a selection of literature concerning the bandwidth choice), or an iteration of the procedure.

It is of practical interest to weaken the assumption of an error process that is independent of the explanatory variables. However, an efficiency gain without any parametric assumptions on the covariance between $(X_t)_{t\in\mathbb{Z}}$ and $(U_t)_{t\in\mathbb{Z}}$ is not expected: the prewhitening method works because the estimator of the filtered error process converges with a faster rate than the local polynomial estimator itself. This is due to the fact that the local polynomial fit uses only data within the local neighborhood, whereas all data points are included in the estimation of the prewhitened error process. If $(X_t)_{t\in\mathbb{Z}}$ and $(U_t)_{t\in\mathbb{Z}}$ are both dependent, we are not able to make use of this mismatch of convergence rates anymore. In this case, the approximation of the error process converges with the same nonparametric rate as the polynomial fit itself. For this reason, an efficiency gain can only be expected if the covariance structure between the covariate and the error process is of parametric nature. In the case that $\mathrm{Cov}(U_s, U_t|X_1, \dots, X_n) = \gamma_{st}(\theta)$ for some parametric error covariance functions $\gamma_{st} \colon \mathbb{R}^d \to \mathbb{R}$, Martins-Filho & Yao (2009) and Su et al. (2013) proposed more efficient estimators based on a prewhitening approach similar to the one derived in this thesis. However, to the authors' knowledge, a more efficient estimation procedure for strongly mixing errors or any other covariance assumption that is equally general has yet not been explored.

In this thesis, we supposed an univariate regression model. When the explanatory variable is multivariate, the introduced filtering method can be applied with some changes in the dimensionality of various quantities. The extension of our asymptotic results to the multivariate regression model as conducted by Xiao et al. (2003) and Su & Ullah (2006) for linear and nonparametric error covariance structures respectively should also be straightforward. However, note that for multivariate data sets, the theoretical asymptotic relative efficiency of the proposed estimator over the conventional one is $(\sigma^2_{\overline{U}}/\sigma^2_U)^{(2p+2)/(2p+2+d)}$, with d the dimension of the explanatory variable. Hence, the relative efficiency gain decreases dramatically for a growing number of explanatory variables. Further, by the *curse of dimensionality* as it was termed by Bellman (1961), much larger datasets than in the univariate case are required to obtain a precise estimation - even for a moderate number of explanatory variables. For such datasets, the runtime of the standard local polynomial estimator is too large to be of practical interest. This problem can be overcome by, e.g. the application of some dimensionality reduction principles (see, e.g. Arenas-Garcia et al. (2013)), which should be kept in mind when extending our findings to the multivariate model.

We now discuss the application of the introduced filtering procedure to other estimation methods. It should be straightforward to apply our method to any other nonparametric estimator that is based on the idea of local (weighted) averaging. The two-step procedure to derive a more efficient estimator for error processes with AR(∞)-correlation was initially formulated for the Nadaraya-Watson estimator[1] (Xiao et al. 2002). The corresponding estimator was shown to be asymptotically normal with a different bias but the same variance expression as the local polynomial estimator. We expect similar asymptotic results if the Gasser-Müller estimator (Gasser & Müller 1979) is used instead, since its asymptotic behavior is similar to that of the local polynomial estimator. In contrast to that, the inclusion of the dependence structure in the estimation procedure of some more recent kernel methods as, e.g. support vector machines (SVM), is an open field. Just recently, Steinwarth & Anghel (2009) were able to show that SVM actually provide consistent estimates if the observation errors are strongly mixing, and Hang & Steinwarth (2014) provided faster learning rates for this setting. But to the author's knowledge, there are no results available yet that address prewhitening techniques for this kind of kernel methods.

Another interesting aspect for future research is the fusion of our method and the kernel signal to noise ratio (KSNR) introduced by Gómez-Chova & Camps-Valls (2012) which accounts for signal and noise relations for nonlinear regression with kernels. The authors applied the KSNR to

[1] See Nadaraya (1964), Watson (1964).

various nonparametric regression models with autocorrelated noise, demonstrating that KSNR reduces the MSE of the conventional kernel ridge regression (KRR) estimator. However, the formulation of the KSNR requires a prior estimation of the noise signal, or a parametric linear model of the noise that needs to be kernelized. Either way, one can conjecture that the filtering method proposed in this thesis might improve the efficiency of the KSNR even further, especially if the data possesses heteroscedastic error terms.

In summary, we think that our method can provide more precise estimations for a wide set of nonparametric time series situations. The asymptotic analysis can hold as a basis for multiple future research concerns.

Notations

Symbols

$\mathbb{R}$	set of all real numbers
$\mathbb{Z}$	set of all integers
$\mathbb{N}$	set of all positive integers
$(Z_t)_{t\in\mathbb{Z}}$	sequence of random variables over some probability space $(\Omega, \mathcal{A}, \mathrm{P})$
$\mathbb{E}(\cdot)$	expectation
$\mathbb{E}(\cdot\vert\cdot)$	conditional expectation
$\mathrm{Var}(\cdot)$	variance
$\mathrm{Cov}(\cdot,\cdot)$	covariance
$1.(\cdot)$	indicator function: $1_A(x) = 1$, if $x \in A$ and $1_A(x) = 0$ otherwise
$\Vert\cdot\Vert_k$	k-norm of a random variable Z: $\Vert Z\Vert_k = \left(\mathbb{E}\vert Z\vert^k\right)^{1/k}$
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$\mathcal{U}[a, b]$	continuous uniform distribution on the interval $[a, b]$
$\mathcal{B}(n, p)$	binomial distribution with number of observations n and success probability p
$\mathbf{A}^\top$	transpose of a matrix $\mathbf{A}$
$[\mathbf{a}]_k$	kth element of a vector $\mathbf{a}$
$[\mathbf{A}]_{k,l}$	lth element in the kth row of a matrix $\mathbf{A}$
$\mathbf{e}_k$	kth unit vector
$\lfloor x\rfloor$	largest integer less than or equal to x
$\sim$	$Z \sim Q$ asserts that the random variable Z is distributed according to the probability distribution Q
$\stackrel{d}{=}$	$Y \stackrel{d}{=} Z$ if Y and Z have the same probability distribution
$\stackrel{\mathrm{P}}{\to}$	convergence in probability
$\stackrel{d}{\to}$	convergence in distribution
$o(a_n)$	Landau notation: $\mathbf{B}_n = o(a_n)$, if $\Vert\mathbf{B}_n\Vert_1/\vert a_n\vert \underset{n\to\infty}{\longrightarrow} 0$
$O(a_n)$	Landau notation: $\mathbf{B}_n = O(a_n)$, if $\limsup_{n\to\infty} \Vert\mathbf{B}_n\Vert_1/\vert a_n\vert < \infty$.
$o_P(a_n)$	Landau notation in probability: $\mathbf{B}_n = o_P(a_n)$, if $\Vert\mathbf{B}_n\Vert_1/\vert a_n\vert \stackrel{\mathrm{P}}{\to} 0$

$O_P(a_n)$	Landau notation in probability: $\mathbf{B}_n = O_P(a_n)$, if for all $\varepsilon > 0$ we can find a constant C_ε such that $\mathrm{P}(\|\mathbf{B}_n\|_1 \geq C_\varepsilon |a_n|) \leq \varepsilon$, for all $n \in \mathbb{N}$
$f_{Z_{t_1},\ldots,Z_{t_k}}$	probability density function of the random vector $(Z_{t_1}, \ldots, Z_{t_k})$
log	natural logarithm
$\alpha(\cdot)$	strong mixing coefficient (see Definition 2.1)
s_X	sample standard deviation
C	generic positive finite constant that may change its value even within a single calculation

Abbreviations

e.g.	for example
i.e.	that is
i.i.d.	independent and identically distributed
IMSE	integrated mean squared error
KRR	kernel ridge regression
KSNR	kernel signal to noise ratio
LASSO	least absolute shrinkage and selection operator
MSE	mean squared error
NLEP	nonlinear error process
RE	relative efficiency
SVM	support vector machine

Bibliography

Aitken, A. C. (1934), 'On least squares and linear combinations of observations', *Proceedings of the Royal Society of Edinburgh* **55**, 42 – 48.

Andrews, D. W. K. (1984), 'Non-strong mixing autoregressive processes', *Journal of Applied Probability* **21**, 930–934.

Arenas-Garcia, J., Petersen, K., Camps-Valls, G. & Hansen, L. (2013), 'Kernel multivariate analysis framework for supervised subspace learning: a tutorial on linear and kernel multivariate methods', *IEEE Signal Processing Magazine* **30**, 16–29.

Bauer, F. L. & Fike, C. (1960), *Numerische Mathematik 2*, Springer.

Bellman, R. E. (1961), *Adaptive control processes*, Princeton University Press.

Billingsley, P. (1969), *Convergence of probability measures*, John Whiley and Sons.

Billingsley, P. (1986), *Probability and measure*, Wiley.

Blumenthal, L. M. (1926), 'Concerning the remainder term in taylor's formula', *American Mathematics Monthly Journal* **33**, 424–426.

Bradley, R. C. (1983), 'Approximation theorems for strongly mixing random variables', *Michigan Mathematics Journal* **30**, 69–81.

Bradley, R. C. (2007), *Introduction to strong mixing conditions*, Kendrick Press.

Brem, R. & Kruglyak, L. (2005), 'The landscape of gene complexity across 5.700 gene expression traits in yeast', *Proceedings of the National Academy of Sciences* **102**, 1572–1577.

Brockwell, P. J. & Davis, R. A. (1990), *Time series: theory and methods*, Springer Series in Statistics.

Caroll, R. J., Lin, X., Linton, O. B. & Mammen, E. (2004), Accounting for correlation in marginal longitudinal regression, *in* D. Y. Lin & P. J. Heagerty, eds, 'Second Seattle Symposium on Biostatistics. Lecture Notes in Statistics.', Springer.

Caroll, R. J. & Ruppert, D. (1988), *Transformation and weighting in regression*, Chapman and Hall.

Cleveland, W. S. (1979), 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association* **74**, 829–836.

Cramér, H. & Wold, H. (1936), 'Some theorems on distribution functions', *Journal of the London Mathematical Society* **11**, 290–294.

Daye, J. & Chen, J. (2012), 'High-dimensional heteroscedastic regression with an application to eqtl data analysis', *Biometrics* **68**, 316–326.

Doukhan, P. (1994), *Mixing. Properties and examples*, Springer.

Engle, R., Granger, W., Rice, J. & Weiss, A. (1986), 'Semiparametric estimates of the relation between weather and electricity sales', *Journal of the American Statistical Association* **81**, 310–320.

Fan, J. (1992), 'Design-adaptive nonparametric regression', *Journal of the American Statistical Association* **87**, 998–1004.

Fan, J. (1993), 'Local linear regression smoothers and their minimax efficiency', *Annals of Statistics* **21**, 196–216.

Fan, J., Gasser, T., Gijbels, I., Brockmann, M. & Engel, J. (1997), 'Local polynomial regression: optimal kernels and asymptotic minimax efficiency', *Annals of the Institute of Statistical Mathematics* **49**, 79–99.

Fan, J. & Gijbels, I. (1992), 'Variable bandwidth and local linear regression smoothers', *Annals of Statistics* **20**, 2008–2036.

Fan, J. & Gijbels, I. (1995), 'Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation', *Journal of the Royal Statistical Society* **57**, 371–394.

Fan, J. & Gijbels, I. (1996), *Local polynomial modelling and its applications*, Chapman and Hall.

Fan, J. & Jiang, J. (2000), 'Variable bandwidth and one-step local m-estimator', *Science in China Series A: Mathematics* **43**, 65–81.

Fan, J. & Yao, Q. (1998), 'Efficient estimation of conditional variance functions in stochastic regression', *Biometrika* **85**, 645–660.

Ferenstein, E. & Gaşowski, M. (2004), 'Modelling stock returns with ar-garch processes', *Statistics and Operations Research Transactions* **28**, 55–68.

Gasser, T. & Müller, H.-G. (1979), Kernel estimation of regression functions, *in* 'Smoothing Techniques for Curve Estimation. Lecture Notes in Mathematics.', Vol. 757, Springer, pp. 23–68.

Gasser, T., Sroka, L. & Jennen-Steinmetz, C. (1986), 'Residual variance and residual pattern in nonlinear regression', *Biometrika* **73**, 625–633.

Gijbels, I. & Fan, J. (1995), 'Adaptive order polynomial fitting: bandwidth robustification and bias reduction', *Journal of Computational and Graphical Statistics* **4**, 213–227.

Gómez-Chova, L. & Camps-Valls, G. (2012), Learning with the signal to noise ratio, *in* 'IEEE International Workshop on Machine Learning for Signal Processing'.

Greene, W. H. (1993), *Econometric Analysis*, Prentice Hall.

Hall, P. & Caroll, R. J. (1989), 'Variance function estimation in regression: the effect of estimating the mean', *Journal of the Royal Statistical Society* **51**, 3–14.

Hang, H. & Steinwarth, I. (2014), 'Fast learning from a-mixing observations', *Journal of Multivariate Analysis* **127**, 184–199.

Härdle, W. & Tsybakov, A. (1997), 'Local polynomial estimators of the volatility function in nonparametric autoregression', *Journal of Econometrics* **81**, 233–242.

Hart, J. D. (1996), 'Some automated methods of smoothing time-dependent data', *Journal of Nonparametric Statistics* **6**, 115–142.

Harvey, A. & Koopman, S. (1993), 'Forecasting hourly electricity demand using time-varying splines', *Journal of the Royal Statistical Society* **88**, 1228–1236.

Hastie, T. J. & Tibshirani, R. (1990), *Generalized additive models*, Chapman and Hall.

Jin, S., Su, L. & Xiao, Z. (2015), 'Adaptive nonparametirc regression with conditional heteroskedasticity', *Econometric Theory* **31**, 1153–1191.

Lim, C., Sen, T. & Peddada, K. (2010), 'Statistical inference in nonlinear regression under heteroscedasticity', *Sankhya B.* **72**, 202–218.

Lin, X. & Caroll, R. J. (2000), 'Nonparametric function estimation for clustered data when the predictor is measured without/with error', *Journal of the American Statistical Association* **95**, 520–534.

Linton, O. B. & Mammen, E. (2008), 'Nonparametric transformation to white noise', *Journal of Econometrics* **142**, 241–264.

Linton, O. B. & Xiao, Z. (2007), 'A nonparametric regression estimator that adapts to error distribution of unknown form', *Econometric Theory* **23**, 371–413.

Maeshiro, A. (1996), 'Teaching regressions with a lagged dependent variable and autocorrelated disturbances', *The Journal of Economic Education* **27**, 72–84.

Marno, V. (2004), *A guide to modern econometrics*, John Whiley and Sons.

Martins-Filho, C. & Yao, F. (2009), 'Nonparametric regression estimation with general parametric error covariance: a more efficient two-step estimator', *Journal of Multivariate Analysis* **100**, 309–333.

Masry, E. (1996*a*), 'Multivariate local polynomial regression estimation: uniform strong consistency and rates', *Time Series Analysis* **17**, 571–599.

Masry, E. (1996*b*), 'Multivariate regression estimation: local polynomial fitting for time series', *Stochastic Processes and their Applications* **65**, 81–101.

Masry, E. & Fan, J. (1997), 'Local polynomial estimation of regression functions for mixing processes', *Scandinavian Journal of Statistics* **24**, 165–179.

Masry, E. & Tjøstheim, D. (1995), 'Nonparametric estimation and identification of non-linear arch time series', *Econometric Theory* **11**, 258–289.

Masry, E. & Tjøstheim, D. (1997), 'Additive non-linear arx time series and projection estimates', *Econometric Theory* **13**, 214–252.

Merlevède, F., Peligrad, M. & Rio, E. (2009), 'Bernstein inequality and moderate deviations under strong mixing conditions', *High Dimensional Probability* **5**, 273–292.

Mokkadem, A. (1988), 'Mixing properties of arma processes', *Stochastic Processes and their Applications* **29**, 309–315.

Nadaraya, E. A. (1964), 'On estimating regression', *Theory of Probability and Its Applications* **9**, 141–142.

Neumann, M. H. (1994), 'Fully data-driven nonparametric variance estimators', *Statistics* **25**, 189–212.

Priestley, M. B. (1981), *Spectral Analysis and Time Series*, Academice Press.

Rosenblatt, M. (1956), 'A central limit theorem and a strong mixing condition', *Proceedings of the National Academy of Sciences* pp. 43–47.

Ruckstuhl, A. F., Welsh, A. H. & Caroll, R. J. (2000), 'Nonparametric function estimation of the relationship between two repeatedly measured variables', *Statistica Sinica* **10**, 51–71.

Ruppert, D. (1997), 'Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation', *Journal of the American Statistical Association* **439**, 1049 – 1062.

Ruppert, D., Sheather, S. J. & Wand, M. P. (1995), 'An effective bandwidth selector for local least squares regression', *Journal of the American Statistical Association* **90**, 1257–1270.

Ruppert, D. & Wand, M. P. (1994), 'Multivariate locally weighted least squares regression', *Annals of Statistics* **22**, 1346–1370.

Ruppert, D., Wand, M. P., Hölst, U. & Hössjer, O. (1997), 'Local polynomial variance function estimation', *Technometrics* **39**, 262 – 273.

Seifert, B., Brockmann, M., Engel, J. & Gasser, T. (1994), 'Fast algorithms for nonparametric curve estimation', *Journal of Computational and Graphical Statistics* **3**, 192–213.

Seifert, B. & Gasser, T. (1996), 'Finite-sample variance of local polynomials: analysis and solutions', *Journal of the American Statistical Association* **91**, 267–275.

Seifert, B. & Gasser, T. (2000), 'Data adaptive ridging in local polynomial regression', *Journal of Computational and Graphical Statistics* **9**, 338–360.

Severini, T. A. & Staniswalis, J. G. (1994), 'Quasi-likelihood estimation in semiparametric models', *Journal of the American Statistical Association* **89**, 501–511.

Shao, J. (2003), *Mathematical statistics*, Springer.

Silverman, B. W. (1986), *Density estimation for statistics and data analysis*, Chapman and Hall.

Spivak, M. (2006), *Calculus*, Publish or Perish.

Steinwarth, I. & Anghel, M. (2009), 'Consistency of support vector machines for forecasting the evolution of an unknown ergodic dynamical system from observations with unknown noise', *Annals of Statistics* **37**, 841–875.

Stone, C. J. (1977), 'Consistent nonparametric regression', *Annals of Statistics* **5**, 595–645.

Su, L. & Ullah, A. (2006), 'More efficient local polynomial estimation in nonparametric regression with nonparametric autocorrelated errors', *Econometric Theory* **22**, 98–126.

Su, L., Ullah, A. & Wang, Y. (2013), 'Nonparametric regression estimation with general parametric error covariance: a more efficient two-step estimator', *Empirical Economics* **45**, 1009–1024.

Su, L., Zhao, Y. & Yan, T. (2012), 'Two-stage method based on local polynomial fitting for a linear heteroscedastic regression model and its application in economics', *Discrete Dynamics in Nature and Society* **2012**, 1–17.

Tsybakov, A. (1986), 'Robust reconstruction of functions by the local-approximation method', *Problemy Peredachi Informatsii* **22**, 133–146.

Vidaurre Henche, D., Bielza, C. & Larrañaga Múgica, P. (2012), 'Lazy lasso for local regression', *Computational Statistics* **27**, 531–550.

Vilar-Fernández, J. M. & Francisco-Fernández, M. (2002), 'Local polynomial regression smoothers with ar-error structure', *TEST* **11**, 439–464.

Vilar-Fernández, J. M. & Francisco-Fernández, M. (2006), 'Nonparametric estimation of the conditional variance function with correlated errors', *Journal of Nonparametric Statistics* **18**, 375–391.

Wand, M. P. & Jones, M. (1995), *Kernel smoothing*, Chapman and Hall.

Watson, G. S. (1964), 'Smooth regression analysis', *Sankhyā: The Indian Journal of Statistics* **26**, 359–372.

Welsh, A. H. (1996), 'Robust estimation of smooth regression and spread functions and their derivatives', *Statistica Sinica* **6**, 347–366.

Wheeden, R. L. & Zygmund, A. (1977), *Measure and integral*, Marcel Dekker.

Wild, C. J. & Yee, T. W. (1996), 'Additive extensions to generalized estimating equation methods', *Journal of the Royal Statistical Society* **58**, 711–725.

Xiao, Z., Linton, O. B., Caroll, R. J. & Mammen, E. (2002), 'More efficient kernel estimation in nonparametric regression with autocorrelated errors', *Journal of Econometrics (2010)* **154**, 186–202.

Xiao, Z., Linton, O. B., Caroll, R. J. & Mammen, E. (2003), 'More efficient local polynomial estimation in nonparametric regression with autocorrelated errors', *Journal of the American Statistical Association* **98**, 980–992.

Zhang, Z. G., Chan, S. C., Ho, K. L. & Ho, K. C. (2008), 'On bandwidth selection in local polynomial regression analysis and its application to multi-resolution analysis of non-uniform data', *Signal Processing Systems Journal* **52**, 263–280.

Ziegelmann, F. (2002), 'Nonparametric estimation of volatility functions: the local exponential estimator', *Econometric Theory* **18**, 985–991.

Ehrenwörtliche Erklärung

Hiermit erkläre ich,

- dass mir die geltende Promotionsordnung der Fakultät für Mathematik und Informatik an der Friedrich-Schiller-Universität Jena bekannt ist,
- dass ich die vorliegende Dissertation selbst angefertigt habe, keine Textabschnitte oder Ergebnisse eines Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen wurden und alle von mir benutzten Hilfsmittel, persönlichen Mitteilungen und Quellen in meiner Arbeit angegeben sind,
- dass alle Personen, die mich bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts unterstützt haben, in der Danksagung erwähnt werden,
- dass die Hilfe eines Promotionsberaters nicht in Anspruch genommen wurde und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorliegenden Dissertation stehen,
- dass ich die vorliegende Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe und
- dass ich weder die gleiche, noch eine in wesentlichen Teilen ähnliche oder eine andere Abhandlung bei einer anderen Hochschule als Dissertation eingereicht habe.

Jena, 22. Juli 2016 Juliane Geller